Depression symptoms across settings: Development and validation of the International Depression Symptom Scale

by

Emily Edmunds Haroz

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

June, 2015

OVERALL ABSTRACT

Background: Existing measurement instruments for depression are most often based on symptoms observed in western clinical populations. It remains unclear whether these instruments are appropriate for use in epidemiologic, screening, intervention monitoring and evaluation, and clinical settings in non-western contexts. The overall goal of this study was to determine if there is a need for new instruments with global applicability to measure depression, and if so, to develop and test this new instrument. **Methods:** Two approaches were used in this process: 1) a systematic literature review of qualitative studies to identify common symptoms related to depression across populations; and 2) a quantitative analysis of existing datasets using item response theory (IRT) to identify how different symptom questions related to depression perform across settings. Results from these investigations were used to inform the development of the International Depression Symptom Scale (IDSS), an instrument designed to reflect global presentations of depression. The IDSS was tested in a community sample of adults (N = 147) in Yangon, Myanmar. **Results:** Results from the literature review and quantitative analysis indicated that most symptoms included in western definitions of depression and on existing measurement instruments are frequently mentioned and perform well across settings. However, additional symptoms need to be included in measurement instruments to more accurately reflect the presentation of depression all over the world. These include: *social isolation, anger, hopelessness, thinking too much, confusion,* and *somatic complaints*. The IDSS was developed based on these conclusions. Testing results showed that the IDSS had high internal consistency reliability (α = 0.92), test-retest reliability (r = 0.87) and inter-rater reliability (ICC = 0.90). Construct, criterion, and incremental validity were also supported for the IDSS. Preliminary evidence supports the IDSS use as a screening tool to detect depressive disorders and impaired functioning in this context. Further research needs to be done to explore its validity in other settings and its use as a clinical or epidemiologic tool. **Conclusions:** Findings contribute to our understanding of how depression manifests globally and demonstrates initial evidence to support

the usefulness of a measurement instrument created to reflect the global presentation of depression.

DISSERTATION COMMITTEE MEMBERS

Judith Bass, Ph. D. (Advisor)

William Eaton, Ph. D.

Kitty Chan, Ph. D. (Chair)

Alden Gross, Ph. D.

DISSERTATION COMMITTEE MEMBER ALTERNATES

Joseph Gallo, M.D.

Courtland Robinson, Ph. D.

ACKNOWLEDGEMENTS

This dissertation is a culmination of my work as a doctoral candidate in the Department of Mental Health at Johns Hopkins Bloomberg School of Public Health. It reflects the generous support, guidance, input and collaboration I have received from multiple mentors, colleagues, family and friends while at Hopkins. I count being a student in the Department of Mental Health as one of the great privileges in my life, and I owe a great debt of gratitude to the department's members and the lessons I have learned during my time here.

First, I owe my upmost appreciation and gratitude to my advisor, Dr. Judith Bass. I have had the benefit of working with Judy for the past five years, starting as a Master's student and through my work as a doctoral student. Over the course of this time, I have been constantly impressed by her intellect and approach to research, which have been fundamental in shaping the field of Global Mental Health. However, it is her deep commitment to mentoring that has helped me the most. She has spent countless hours helping to guide my own research interests, discuss rigorous methods, and work to effectively communicate the process and results. As I continue my career in public health, I will constantly strive to meet her high standards and expectations.

I have also had the great privilege of having Dr. Paul Bolton as a mentor throughout my time at Hopkins. I originally met Paul at a conference while doing my master's work at Columbia University. After listening to his talk about a project in Uganda, I felt like I had finally found exactly what I wanted to do with my life. About four years ago, Paul asked me whether I had any ideas for what I wanted to write my dissertation on. During our discussion, he drew a Venn diagram, which has served as the inspiration for this whole dissertation project. None, of this would have been possible, without the vote of confidence, constant encouragement, and thoughtful reflections, Paul has provided over the past four years. I feel incredibly lucky to have the benefit of his mentorship during my time at Hopkins.

I wish to thank my training grant director, Dr. Peter Zandi, who has done an incredible job of instilling in his mentees an ability to critically think about psychiatric epidemiology. Some

of the most rich and engaging discussions about methodology during my time at Hopkins have come from our monthly journal club meetings. I also would like to acknowledge the National Institute of Mental Health, which generously funded my position on the Psychiatric Epidemiology Training Grant.

I wish to thank my dissertation committee readers, Dr. Kitty Chan, Dr. William Eaton, and Dr. Alden Gross, who have already provided me with valuable insights and suggestions related to measurement of mental disorders. Your contributions have made me a better researcher, while your rich insights have helped to contribute to improving the discussion.

Over the course of my time at Hopkins, I have had the honor to work with whole MHAP and ATIP team in Thailand and Myanmar. I wish to thank all of them for their constant commitment to improving mental health and conducting high quality research even in less than ideal circumstances. Dr. San San Oo, Dr. Win Naing, Dr. Kyaw Lin, and Dr. Thapyay Kyi, have shown an incredible dedication to improving the mental health of people in Myanmar. I have been particularly inspired through my work with the MHAP project's two chief clinical supervisors, Kyaw So Win and Saw Tet Tun, who are very talented and true inspirations of how to overcome incredible adversity to reach ones goals. None of the research would be possible without Lae Lae Nwin and Yasmina Yules. I feel very thankful to work with these two women, and even more thankful to have them both as friends. Working with Dr. Catherine Lee, has been a true privilege. Cate is fully dedicated to improving the health and wellbeing of people from Burma. It has been a great honor to call her a colleague and a dear friend.

I also want to extend my deepest gratitude to my friends and family. To my friends in Baltimore and elsewhere, thank you for providing much of the work-life balance. It has been a joyful couple of years because of all of you. A special thank you to BJ who is always is there to make me laugh.

To my in-laws, Vicki Jackson, Robert Taylor, Michael Taylor, and Sophie Taylor, I greatly appreciate your constant love, support and understanding. I am very lucky to be included

in your family. To my brother and sister-in-law, David Haroz and Elissa Margolin, thank you for being curious about what I do. Your words of encouragement throughout this whole process have been incredibly helpful. To my brother Carim and sister Jamie, thank you for all the love and support over the past five years.

To my parents, words cannot fully express how much your unflappable confidence and unwavering support has meant to me. I could not have done this without you. From the countless hours you have spent helping in Baltimore, Vashon or Cambridge, to the feedback you have given on every paper, to the numerous phone conversations providing guidance or just a good listening ear, you have encouraged and helped me pursue my passions and dreams. You have always believed in me and my abilities, even when I was not as sure. For this, I am forever grateful.

Last, but certainly not least, I am truly grateful to my husband, Jake and my daughter Madeline. I knew that coming to Hopkins would lead to great professional satisfaction, but I had no idea that I would meet my life partner and best friend here as well. Jake, you are an incredible source of love, support and encouragement. Thank you for helping me to reach my goals and loving me every step of the way. You always hold true to your vows. To Madeline, you are my inspiration. I hope the work I do in the future will help to make the world a better place for you and your generation. Jake and Madeline, I love you.

Intended to be left blank

Contents

List of Tables

List of Figures

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 1. Introduction

1.1 Statement of problem

Major Depressive Disorder is a significant contributor to the global burden of disease. In 2010, it was estimated that the global point prevalence for Major Depressive Disorder was 4.7% (Ferrari et al., 2012). In Low and Middle Income Countries (LMIC) Major Depressive Disorder ranges from being the 4th leading cause of disability (measured in Disability Adjusted Life Years) in Andean Latin America, South East Asia, and North Africa and the Middle East, to being the 19th leading cause in Western Sub-Saharan Africa (Murray et al., 2013). It is the leading cause of Years Lived with Disability (YLD) in many low-income countries (Murray et al., 2013).

What is not clear from these estimates is the reason behind the significant variability in these numbers across settings and populations. Heterogeneity in prevalence estimates of depression has been found in numerous cross-national studies (Moussavi et al., 2007; Weissman et al., 1996), and is even more pronounced in the context of trauma-affected populations. Estimates among populations affected by armed conflict or displacement have been found to range from 3% to 85.5% (Steel et al., 2009). Variation may reflect true differences in the epidemiology of depression, or it may be a result of artificial differences caused by measurement error. Measurement error impacts policy level decisions, program planning and evaluation, as well as clinical service provision and decision-making, particularly in low-resource settings (Kohrt et al., 2011; Wessells, 2009).

The complexities involved in measuring depression in different settings may contribute to measurement error related to prevalence estimates. Measurement of depression symptoms in non-western settings has typically taken several approaches. The first, a universalist approach, utilizes western screening questionnaires to assess depressive symptoms. Examples of this approach are the world mental health surveys or global burden of disease studies (Moussavi et al., 2007; Murray et al., 2013). By contrast, a strictly particularlist approach aims to understand locally

relevant syndromes in one setting, and develop measurement tools specific to this population (Miller et al., 2006; Patel, Simunyu, Gwanzura, Lewis, & Mann, 1997; Phan, Steel, & Silove, 2004). The constellation of symptoms that together make up a syndrome, may vary by culture and context. Finally, a third approach has been to combine the two previous approaches by adapting a western based depression symptom screener that includes several locally relevant symptoms in addition to the standard items (see examples Bass, Ryder, Lammers, Mukaba, & Bolton, 2008; Betancourt et al., 2009; Bolton, 2001; Haroz et al., 2014; Rasmussen et al., 2014).

However, there are problems with these approaches. Using western measures that were developed among clinical populations in western, high-income settings with existent mental health care systems can be problematic (Bass, Bolton, & Murray, 2007; Wessells, 2009). Such measures may lack content validity in other cultural settings. Qualitative and ethnographic approaches are less biased toward western culture, but in turn lack generalizability and comparability. The third approach, adapting existing measures through qualitative work, allows for comparison of scores across settings while also ensuring local relevance. However, this approach can also be quite resource intensive (Hollifield, 2002), a particular problem for low-resource settings.

To improve the quality of measurement of depression across settings, there is a need for a more robust and relevant screening measure of depressive symptoms that approaches universal validity while being less biased toward symptoms identified only in western populations. The overall goal of this study is to develop such a measure to be used in public health and clinical settings.

1.2 Public Health Significance

Utilizing a combination of a systematic literature review of qualitative studies related to depression symptomology and an Item Response Theory (IRT) analysis of data from eight diverse settings, this study aims to facilitate the development and testing of a single scale that can

be validly used to measure depressive symptoms in a range of diverse settings. Accurate measurement of depression has public health, research, policy and economic implications. Developing valid and reliable measurement tools has been identified as one of the grand challenges for the field of global mental health (Collins et al., 2011).

Sound measurement serves as the basis for accurately identifying populations with the greatest disease burden. Identifying which, if any, measurement factors contribute to existing differences in prevalence and incidence of mental disorders between countries and ethnic groups is important (Collins et al., 2011). In the last 30 years, the field of psychiatric epidemiology has seen many improvements in cross-population comparison studies, including the use of similar study designs and sampling procedures. These improvements have led to an enhanced ability to compare epidemiological estimates across contexts. However, measurement challenges persist, including response bias in self-report scales and differing interpretations of survey instruments in different settings (Kessler, 2000).

Accurate measurement is also necessary for identifying those within a population who would likely benefit from services. This is particularly true in many LMIC where there are few trained mental health professionals. Non-specialist workers are often the ones used to screen individuals, relying on self-report instruments to inform their decisions. With a growing call for integrating mental health screening and care into primary health care settings and other task-sharing approaches in LMIC (Collins et al., 2011; Patel, Simon, Chowdhary, Kaaya, & Araya, 2009), suitable screening tools that can be easily administered by non-specialist workers are needed to identify those in need and guide decisions about appropriate treatment.

Accurately measuring alleviation or aggravation of illness in individuals and within populations is important for clinical and research purposes. Clinically, measurement instruments can be used for determining how treatment should be tailored and whether someone is responding to an intervention. Recent research has shown that, in the context of a psychotherapeutic intervention for common mental health disorders delivered by non-professionals to trauma-

affected adults, tailoring of treatment and monitoring of patient symptoms can be done by non-specialist workers using self-report instruments (Bolton et al., 2014; L. K. Murray et al., 2013). Without valid and reliable instruments inappropriate treatment decisions are possible and potentially harmful (Wessells, 2009). Research into what treatments work, how they work, and for which populations, depends on accurate measurement tools as well. Without such tools, it is impossible to know whether changes in symptoms is real, or is simply an artifact due to measurement error.

Given existing challenges with measuring depression in LMIC and different settings, it is important to overcome some of the limitations of existing instruments. Programs that provide psychiatric care may not always have the time or resources to generate a culturally appropriate tool for measuring depression symptoms in their particular setting. A scale that is based on empirical evidence of the presentation of depression globally may be particularly useful for these situations. Moreover, a scale that is based on global presentations of depression, rather than western clinical presentations of depression, has the potential to more accurately reflect depression across a range of settings, making epidemiologic estimates more comparable.

1.3 Specific Aims

The present study will examine depression symptomology across a wide range of settings and test a self-report measure created based on these findings. The specific aims of this dissertation are:

Aim 1. To identify a set of signs and symptoms which have been described in ethnographic, anthropological, and qualitative research on depressive-like syndromes in a range of cultures and settings.

Aim 2. To quantitatively identify signs and symptoms from a commonly used western-developed adult depression screening measures that are applicable and unbiased across multiple diverse settings.

Aim 3. To test in a low-resource setting the reliability, validity, and clinical utility of a self-report measure developed based on the evidence from Aims 1 & 2.

1.4 Overview of dissertation

Chapter 2 presents a brief overview of the history of identification and diagnoses of, the evolution of the field cross-cultural mental health, and the use of psychometric evaluation in the context of research on psychiatric epidemiology and global mental health, in order to orient the reader to issues that have helped guide this investigation. Chapter 3 is an overall methods chapter that provides information on the analytic methods used throughout this dissertation. These methods include a systematic review of qualitative studies related to depression, an IRT analysis of data from eight diverse settings, and reliability and validity testing of an instrument to measure depression. Chapter 4 is the first results chapter and refers to Specific Aim 1. This chapter presents the results from the systematic review of symptoms of depression that are mentioned in qualitative studies from across the world. This review provides a comprehensive picture of what are common symptoms of depression globally and what may be missing from current depression screening measures. Chapter 5 presents the results of the analyses for Specific Aim 2, focusing on the quantitative analyses utilizing IRT to examine the performance of symptoms from a commonly used measure of depression across eight diverse settings. The goal of the quantitative analysis is to look at which of the symptoms currently used in a common depression measure are most informative across different populations. The evidence generated from Chapters 4 and 5 was then used to identify a set of signs and symptoms that are representative of depression and demonstrably less biased in multiple settings. This evidence served as the foundation for development of a new instrument to measure depression globally—the International Depression

Symptom Scale (IDSS). Chapter 6, related to Specific Aim 3, presents initial evidence of the reliability, validity, and clinical utility of the IDSS from testing done in Yangon, Myanmar. The instrument testing study was carried out in January 2015, with support from the United States Agency for Instrument Development (USAID) Victims of Torture Fund (VOT). The dissertation was completed with support from the National Institute of Mental Health (NIMH) T32MH014592-38.

References

Bass, J. K., Bolton, P. A., & Murray, L. K. (2007). Do not forget culture when studying mental health. *The Lancet, 370*(9591), 918-919.

Bass, J. K., Ryder, R. W., Lammers, M., Mukaba, T. N., & Bolton, P. A. (2008). Post-partum depression in kinshasa, democratic republic of congo: Validation of a concept using a mixed-methods cross-cultural approach. *Tropical Medicine & International Health, 13*(12), 1534-1542.

Betancourt, T. S., Bass, J., Borisova, I., Neugebauer, R., Speelman, L., Onyango, G., & Bolton, P. (2009). Assessing local instrument reliability and validity: A field-based example from northern Uganda. *Social Psychiatry and Psychiatric Epidemiology, 44*(8), 685-692.

Bolton, P., Lee, C., Haroz, E. E., Murray, L., Dorsey, S., Robinson, C., . . . Bass, J. (2014). A transdiagnostic community-based mental health treatment for comorbid disorders: Development and outcomes of a randomized controlled trial among Burmese refugees in Thailand. *PLoS Medicine, 11*(11), e1001757.

Bolton, P. (2001). Cross-cultural validity and reliability testing of a standard psychiatric assessment instrument without a gold standard. *The Journal of Nervous and Mental Disease, 189*(4), 238-242.

Collins, P. Y., Patel, V., Joestl, S. S., March, D., Insel, T. R., Daar, A. S., . . . Fairburn, C. (2011). Grand challenges in global mental health. *Nature, 475*(7354), 27-30.

Ferrari, A., Somerville, A., Baxter, A., Norman, R., Patten, S., Vos, T., & Whiteford, H. (2012). Global variation in the prevalence and incidence of major depressive disorder: A systematic review of the epidemiological literature. *Psychological Medicine, 43*(3), 471-481.

Haroz, E. E., Bass, J. K., Lee, C., Murray, L. K., Robinson, C., & Bolton, P. (2014). Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the Thailand Burma border. *BMC Psychology, 2*(1), 31-40.

Hollifield, M. (2002). Accurate measure in cultural psychiatry: Will we pay the costs? *Transcultural Psychiatry, 39*, 419–421.

Kessler, R. C. (2000). Psychiatric epidemiology: Selected recent advances and future directions. *Bulletin of the World Health Organization, 78*(4), 464-474.

Kohrt, B. A., Jordans, M. J. D., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). Validation of cross-cultural child mental health and psychosocial research instruments: Adapting the depression self-rating scale and child PTSD symptom scale in nepal. *BMC Psychiatry, 11-27.*

Miller, K. E., Omidian, P., Quraishy, A. S., Quraishy, N., Nasiry, M. N., Nasiry, S., . . . Yaqubi, A. A. (2006). The Afghan symptom checklist: A culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry, 76*(4), 423-433.

Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., & Ustun, B. (2007). Depression, chronic diseases, and decrements in health: Results from the world health surveys. *The Lancet, 370*(9590), 851-858.

Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., . . . Abdalla, S. (2013). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010. *The Lancet, 380*(9859), 2197-2223.

Murray, L. K., Dorsey, S., Haroz, E., Lee, C., Alsiary, M. M., Haydary, A., . . . Bolton, P. (2013). A common elements treatment approach for adult mental health problems in low-and middle-income countries. *Cognitive and Behavioral Practice, 21*(2), 111-123.

Patel, V., Simon, G., Chowdhary, N., Kaaya, S., & Araya, R. (2009). Packages of care for depression in low-and middle-income countries. *PLoS Medicine, 6*(10), e1000159.

Patel, V., Simunyu, E., Gwanzura, F., Lewis, G., & Mann, A. (1997). The Shona symptom questionnaire: The development of an indigenous measure of common mental disorders in Harare. *Acta Psychiatrica Scandinavica, 95*(6), 469-475.

Phan, T., Steel, Z., & Silove, D. (2004). An ethnographically derived measure of anxiety, depression and somatization: The Phan Vietnamese psychiatric scale. *Transcultural Psychiatry, 41*(2), 200-232.

Rasmussen, A., Eustache, E., Raviola, G., Kaiser, B., Grelotti, D. J., & Belkin, G. S. (2014). Development and validation of a haitian creole screening instrument for depression. *Transcultural Psychiatry, 52*(1), 33-57.

Steel, Z., Chey, T., Silove, D., Marnane, C., Bryant, R. A. & van Ommeren, M. (2009). Association of torture and other potentially traumatic events with mental health outcomes among populations exposed to mass conflict and displacement: A systematic review and meta-analysis. *JAMA, 302*(5), 537-549.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H. G., . . . Lellouch, J. (1996). Cross-national epidemiology of major depression and bipolar disorder. *JAMA, 276*(4), 293-299.

Wessells, M. G. (2009). Do no harm: Toward contextually appropriate psychosocial support in international emergencies. *The American Psychologist, 64*(8), 842-854.

Chapter 2. Background

2.1 History of identification and diagnoses of depression

The modern concept of depression has its roots in ancient Egypt, Acedia and Melancholia, illustrating that syndromes characterized by grief, sadness, dejection, sorrow and associated alterations in behavior, have long been recognized in humans. Perhaps the first recorded mention of a depressive like illness came from the Ebers Papyrus in Ancient Egypt in 1550 BC. In the text there is mention of a condition "when his heart is afflicted and has tasted sadness, behold his heart is closed in and darkness in in his body because of anger which is eating up his heart" (Ghalioungui, 1987). Syndromes characterized by depressed mood, reduced energy and negative thoughts about one's self, have long been recognized in Western settings as well. Melancholia was mentioned in Hippocratic writings in ancient Greece and was described as a chronic form of madness that was characterized by fearfulness, sadness, fatigue, and occasional gastrointestinal problems, aversion to food, despondency, sleeplessness, irritability, and restlessness (Jackson, 1969; Jones, Withington, & Potter, 1928).

With the growing popularity of Christianity in the 4th and 5th centuries, guilt associated with chronic dejected states, slowly emerged and gave rise to the condition *Acedia*. Acedia was considered one of the cardinal sins of Christianity (Jackson, 1985) and was detailed by John Cassian around the beginning of the 5th century A.D. to describe problematic behavior in monks (Altschule, 1965). Cassian associated Acedia with behavior related to boredom that led the monk to give up his religious devotion and led to anxiety of the heart (Altschule, 1965). Cassian characterized Acedia as a sin and considered any illnesses or abnormal behavior as being attributed to transgressions against God (Altschule, 1965). Acedia has many parallels to modern day depression as it was associated with deep sadness, dejection, sorrow and unexplained or strange behavior (Jackson, 1985).

Both Melancholia and Acedia share common symptoms such as sorrow, sadness, fear, and exhaustion. However, Melancholia was considered a medical disease and sometimes

associated with delusional beliefs, whereas Acedia was considered a sin and was never described as being characterized by delusions (Jackson, 1985). By about the 15th and 16th centuries, as the position of the Christian church began to recede, Acedia started gradually becoming less important and slowly became intertwined with melancholia (Jackson, 1985)

Acedia, Melancholia, and depressive disorders all can be thought of as explanatory models of similar underlying phenomena. While these conditions differ in terms of their presumed origins and societal repercussions, they all reflect underlying distress manifested in manners consistent with the social atmosphere of their times.

Modern day depression is defined by prominent classification systems, such as the Diagnostic and Statistical Manual (DSM-V; American Psychiatric Association, 2013) and International Classification of Disease (ICD-10; World Health Organization, 1992). The DSM is a diagnostic manual produced by the American Psychiatric Association and only contains information about psychiatric disorders, in contrast to the ICD which includes psychiatric disorders and diseases related to physical health. The current version of the DSM classifies all Depressive Disorders under the category of mood disorders, which also include Bipolar Disorders, Mood Disorders Due to General Medical Conditions, and Substance-Induced Mood Disorders. Depressive Disorders consist of Major Depressive Disorder (single episode or recurrent), Dysthymic Disorder, Disruptive Mood Dysregulation Disorder, Premenstrual Dysphoric Disorder, and Depressive Disorder Not Otherwise Specified. The symptoms of Major Depressive Disorder (MDD) are depressed mood, lack of interest or pleasure, significant weight loss or gain, insomnia or hypersomnia, psychomotor agitation or slowing, fatigue or loss of energy, feelings of worthlessness or guilt, trouble concentrating and recurrent thoughts or death or suicide ideation (American Psychiatric Association, 2013). To meet full criteria for a major depressive episode, a person must report at least 5 out of the DSM 9 symptoms, including depressed mood or loss of interest, for 2 weeks or longer.

The ICD-10 has a slightly different diagnostic conceptualization of Major Depressive Disorder than the DSM. The 10 symptoms included in the ICD-10 for depression are: 1) persistent sadness or low mood, 2) loss of interests or pleasure, 3) fatigue or low energy, 4) disturbed sleep, 5) poor concentration, 6) low self-confidence, 7) poor or increased appetite, 8) suicidal thoughts or acts, 9) agitation or slowing of movements, 10) guilt or self-blame. Based on ICD-10 criteria, people are classified as "not depressed" (fewer than 2 symptoms), "mild depression" (2 or 3 symptoms), "moderate depression" (4 to 5 symptoms) and "severe depression" (6 or more symptoms, with or without psychotic symptoms) (World Health Organization, 1992).

Both the DSM and the ICD are taxonomies with specific diagnostic criteria for psychiatric illness. More generally, depression is often characterized by a deep sadness and unhappiness that is often accompanied by awareness of pain, lowered mood, and reduced functioning (Bhugra & Bhui, 2007). Other negative and physical symptoms include poor-self attitude, decreased vital sense, and psychomotor vegetative phenomena. It seems that while feeling sad or bad is a natural part of human nature, there is a point when these feelings becomes debilitating, impairing or even-life threatening and subsequently these feelings are elevated to the status of an illness or disorder (Bhugra & Bhui, 2007).

2.2 Brief history of cross-cultural mental health

Cross-cultural comparisons of psychopathology were initially seen as a way to validate mental health related phenomenon observed in the west. While informal cross-cultural research may have taken place prior to the early 1900s, it was Emil Kraepelin's establishment of comparative psychiatry in 1904 that serve as the roots of the modern day discipline of cross-cultural mental health (Jilek, 1995).

Kraepelin traveled to Java (in modern day Indonesia) to examine the ethnic and sociocultural factors of the human mind in health and disease. He believed that human

characteristics were manifested in both religion and customs, and so religion and customs would also be important in expressions of mental disorders. Kraepelin believed that research outside of western society would provide valuable insights into the mental health of other nations and cultures and the results from this research had the potential to contribute to an understanding of universal human psychopathological processes. Kraeplin's goal was to find out "whether certain forms of insanity that form the main content of our institutions, occur in like manner and frequency, as among us, also occur under entirely different conditions of living and among entirely different ethnicities" (Kraepelin, 1904).

While Kraeplin was mainly concerned with finding universals in human behavior, he also found differences in expression and attributed these to differences to the stage of societal development. He noted:

> the relative absence of delusions among the Javanese might be related to lower stage of intellectual development attained and the rarity of auditory hallucinations might reflect the fact that speech counts for far less than it does with us and that thoughts tend to be governed more by sensory images (Kraepelin, 1904).

These attributions demonstrate Kraepelin's understanding of the larger social context of mental illness and how culture and context may shape disease presentation.

Following Kraepelin, H.B.M. Murphy at McGill University and Julian Leff at the Institute of Psychiatry in England, used clinical observations and epidemiological methods in order to "identify, verify and explain the links between mental disorder and these broad psychosocial characteristics [which differentiate nations/people]" (Murphy, 1982). Like Kraepelin, Murphy and Leff believed that mental disorders were related to modernization, however they lacked the social Darwinism perspective that characterized much of Kraepelin's cross-cultural work (Kirmayer, 2007).

By the 1960s and 1970s, Alexander Leighton and Jane Murphy dominated the field of cross-cultural psychiatry with their research in Africa, Alaska and rural Nova Scotia. Both Leighton and Murphy were trained in anthropology and were interested in the impact of social

and cultural influences on individual's mental health (Kirmayer, 2007). Due to their anthropological backgrounds, Leighton and Murphy utilized ethnographic approaches to construct a dimensional understanding of psychopathology within and across cultures. Murphy in her landmark paper in 1976 compared Eskimos in Alaska to individuals in rural Nigeria. In this paper she argues for the universality of psychotic symptoms across different cultural contexts and that distinguishing between individuals who are sane and individuals who are insane, is, in fact, possible in range of diverse settings. (Murphy, 1976)

With the advent of the DSM-III in 1980 (American Psychiatric Association, 1980) the field of psychiatry in the United States had a manual that emphasized the diagnosis of discrete mental disorders, instead of dimensional assessments. DSM-III not only fundamentally changed clinical practice in the United States, but paved the way for more expansive psychiatric epidemiological investigations domestically and globally (Anthony, Eaton, & Henderson, 1995). Structured diagnostic interviews and self-report scales that reflected DSM-III's classification became available, and with them, the ability to evaluate large numbers of people and make comparisons across populations. This led to major cross-national research studies including the International Pilot Study of Schizophrenia (WHO, 1978) and the Cross-National Study of Depression (Sartorius, 1983).

2.3 Global Mental Health

Current cross-cultural mental health work is known as the field of global mental health (GMH). Global mental health has been defined as "the area of study, research and practice that places a priority on improving mental health and achieving equity in mental health for all people wordwide," (Patel & Prince, 2010). GMH is not only concerned with searching for universals, but also provides an avenue for understanding and treating culturally specific problems related to mental health. Cultural concepts or distress and local idioms of distress have been identified in

many contexts across the globe, providing a rich picture of the social and cultural context for psychiatric distress (Kohrt et al., 2014; Patel & Prince, 2010).

Etic vs. emic

Global mental health research has long been centered on the debate between a more etic or more emic approach to the study of mental disorders worldwide. The comparison of western-psychiatric concepts (i.e. DSM-defined disorders) to phenomena in other contexts and the study of culturally-specific disorders illustrates the differences between *etic* and *emic* approaches to psychiatric epidemiology. An etic, or universalist approach, highlights the universality of constructs. Emic, or more particularlist methods, focus on the culturally specific aspects of constructs. While some researchers have advocated for the etic perspective by pointing to the fact that the DSM's constellation of symptoms for certain disorders can be found in many cultures all over the world (Marsella, Friedman, Gerrity, & Scurfield, 1996), others have argued a more emic approach stressing that there is no universal response to stress and manifestation of distress (Summerfield, 2000)

More recently, researchers have used a combination of both etic and emic perspectives to examine mental disorders across cultures (Draguns & Tanaka-Matsumi, 2003; De Jong & Van Ommeren, 2002). This approach emphasizes the universality of the underlying construct (e.g. depression) but understands that expression of these underlying constructs may differ by culture and situation (Maes, Kohrt, & Closser, 2010). This balance between etic and emic recognizes that there may be local signs, symptoms, or syndromes that signify distress, but that often these share similarities with common signs, symptoms and syndromes that exist in other cultures as well. Increasingly, researchers have begun to identify local idioms of distress through qualitative research and used this information to inform decisions on which problems need addressing, adaptation of instruments, clinical intervention selection and adaptation, and monitoring and evaluation of health programs(Bolton, Neugebauer, & Ndogoni, 2002; Bolton, Surkan, Gray, & Desmousseaux, 2012; Bolton, Michalopoulos, Ahmed, Murray, & Bass, 2013; Kaiser et al., 2014;

Rasmussen, Katoni, Keller, & Wilkinson, 2011). The identification of local idioms has the potential to illuminate the locally relevant signs and symptoms of mental disorders allowing for better identification and more culturally relevant treatment (Kohrt & Hruschka, 2010; Nichter, 2010).

2.4 Depression and global mental health

Depression is one of the most studied psychiatric syndromes across cultures, yet it has been a challenge to specify the exact culturally invariant characteristics of the disorder (Draguns & Tanaka-Matsumi, 2003). Many languages do not have a specific word for the syndrome (Bhugra & Bhui, 2007) Changing diagnostic criteria, cultural reactions to depression, and variation in epidemiological methods, have hampered the study of depression across settings and contributed to dramatically different estimates in burden of disease in different populations (Hwu & Compton, 1994).

In an early attempt to overcome these barriers, the World Health Organization (WHO) initiated a major multi-national study in 1983 aimed to systematically measure depression symptomology in fifteen different countries using standardized assessments. The WHO collaborative study (Sartorius, 1983) found that 76% of identified depressed patients reported a common pattern of symptoms that included sadness, absence of joy or pleasure, reduced concentration, lack of energy and a sense of inadequacy. Suicidality was also mentioned in 59% of those identified as depressed. Despite consistent pattern of symptoms, this study found that prevalence rates for depression ranged from 2.6% in Japan to 29.5% in Chile (Sartorius, 1983) Discrepancies such as these may be explained by differing definitions in caseness, variations in diagnostic interview methodologies, differences in environmental circumstances, and cultural differences in symptom expression (Bhugra & Bhui, 2007).

In their review Draguns & Tanaka-Matsumi (2003) identified studies that demonstrated variance in symptom presentation across cultures. In particular, feelings of guilt, have found to

not be associated with depression in Africa India, Indonesia, Japan, and China (Draguns & Tanaka-Matsumi, 2003). Somatization, which is not a focus in western definitions of depression, has been found to be closely related to depression particularly in Chinese populations (Kleinman, 1982). Somatic symptoms have also been found in Japanese, Indian, Latin American, and African samples as well (Kirmayer, 1984). These differences in symptom presentation may contribute to the discrepancies of epidemiologic depression estimates worldwide.

To address culturally variant symptoms of depression some researchers have developed locally relevant screening tools, such as the American Indian Depression Scale (Manson, Shore, & Bloom, 1985), the Shona Symptom Questionnaire (Patel, Simunyu, Gwanzura, Lewis, & Mann, 1997), and the Afghan Symptom Checklist (Miller et al., 2006). These instruments stress the importance of culturally specific idioms of distress which can aid in identifying those most in need of help, evaluating the epidemiological footprint of this distress, and help in the alleviation of suffering. However, because these instruments are so culturally specific, they are limited in their generalizability and do not allow for cross-cultural comparison.

To what extent there are symptoms of depression that are universal and/or symptoms of depression that are contextually specific, still remains to be determined. The cross-national studies by WHO (Sartorius, 1983) used standardized instruments developed in western clinical populations and looked for patterns of symptom endorsement, while others have looked within certain cultures to determine the relevance of particular symptoms. However, few studies have looked at which idioms (outside of Western clinical populations) may be near universal and not directly captured by the current western biomedical psychiatric nosology of depression.

To address part of these limitations researchers have begun to look at how individual items on assessment instruments representing symptoms of depression perform within and across contexts. One approach has been to use Item Response Theory (IRT) to examine Differential Item Functioning (DIF) (Carragher, Mewton, Slade, & Teesson, 2011; H. Choi, Fogg, Lee, & Wu, 2009; Y. Choi, Mericle, & Karachi, 2006; Olino et al., 2012). This approach provides a powerful

tool to examine individual items of a scale to determine which items perform the best across multiple settings. However, research utilizing these methods, especially in cross-cultural mental health studies, is limited.

2.5 Measurement in social science

Similar to the assessment of depression globally, measurement in the social sciences has had a long and important history. Even before science became a disciplinary field, humans used measurement out of necessity to quantify and evaluate social processes (Duncan, 1984). However, it wasn't until a more thorough understanding of probability and statistics developed that the practice of measuring human behavior became popular. Initial interest in quantifying human characteristics and linking these to other outcomes was driven by curiosity in human intelligence. Starting with Sir Francis Galton, Karl Pearson, Alfred Binet, as well as others, mental testing was developed and became popular in the late 19th and early 20th centuries. Mental testing serves as the origin and provides the foundation for of the techniques used in modern day psychometric testing (DeVellis, 2012).

Much of social science is concerned with describing what is not directly observable. For example the theoretical construct of "general intelligence" which can be inferred but not directly observed. These types of constructs have been labeled *latent* variables. In order to quantify latent variables, psychometrics relies on indicators that can be measured through observation. Psychiatric illnesses, when considered in a measurement framework, are treated as latent variables, due to the lack of established biological markers for these disorders. Assessment of psychiatric disorder is then often done through careful, systematic questioning either by a psychiatrist or other highly trained mental health clinician or by a less trained individual using a set of standardized questions which make up a clinical measurement instrument (Murphy, Tsuang, & Tohen, 2002).

Measurement instruments that combine observable indicators into a composite score representing levels of a latent variable are formally called *scales*. Scales are widely used in modern social science and are particularly important in the field of psychiatric epidemiology and public mental health (DeVellis, 2012). Scales generally treat psychopathology as dimensional and are usually used as screening tools. Scales ask about constellations of signs and symptoms that are thought to be interrelated and together form the picture of a disease/disorder. In scoring scales, each item is treated equally and scores are often summed across all items or across pre-specified sub-scales of a limited number of items (i.e. a subscale of somatic symptoms on a depression scale). Because each item is weighted equally, scales do not differentiate between the essential symptoms of a syndrome and other symptoms associated with that syndrome (Murphy et al., 2002).

The other main type of psychiatric measurement instrument is a *schedule*. Schedules are aimed at classifying people into diagnostic categories and are usually used when trying to identify or diagnose cases (case identification and case diagnosis). Schedules are dependent on the concept of a syndrome; an aggregate set of signs and symptoms that together form the picture of a disease. The DSM and ICD are examples of compendiums of syndromes. The main purpose of schedules is to evaluate the presence or absence of a disease/disorder. Most schedules typically begin by asking the respondent about the presence of the cardinal symptoms of a syndrome (e.g. in depression this would be lack of interest and sad affect or mood). If the respondent answers negatively to the initial questions, then the rest of the questions related to other symptoms of the disorder are skipped. In this way, the essential symptoms are weighted differently than other symptoms associated with the syndrome (Murphy et al., 2002).

2.6 Measurement in psychiatric epidemiology

Measurement in psychiatric epidemiology includes methods for case-finding, case identification, case diagnosis, and screening. Case-finding involves locating and identifying

individuals who are potential cases, and can be done through either surveillance or survey methods. Surveillance methods utilize existing records from facilities that treat the disorder under study, while survey methods aim to define a certain target population and measure the number of people within this population who have the disorder. Case identification can be defined as the process of accurately and reliably identifying cases and non-cases from an eligible population. Case diagnosis is the process of determining whether a particular person's signs and symptoms indicate the presence of an established and recognized mental disorder. Screening, can be used as a method in case finding, and refers to the process of identifying individuals in a population who may be at risk or have the disorder, but do not yet have an actual clinical diagnosis (Murphy et al., 2002).

Regardless of the purpose, one method often used in psychiatric epidemiology is the structured diagnostic interview, such as the Diagnostic Interview Schedule (DIS) (Robins, Helzer, Croughan, & Ratcliff, 1981) or the Composite International Diagnostic Interview (CIDI) (Robins et al., 1988), which are schedules that allow lay interviewers or low-level clinicians to make diagnostic classifications (Kessler et al., 1994; Regier et al., 1984). Another method is the use of medical records of psychiatric patients, which have the benefit of being records of psychiatric evaluations and diagnoses, but these are often hard to obtain and limit the study of psychiatric epidemiology to only those who are receiving treatment. Finally, self-report scales used as screening measures are easy to administer and have the flexibility to measure a range of related constructs (e.g. coping, social support, etc.), but are often are less sensitive and specific than schedules or psychiatric evaluations.

Despite their limitations, self-report scales are important in psychiatric epidemiology and public mental health, particularly in global mental health research where the number of trained mental health professionals is limited (Bruckner et al., 2011; Saxena, Thornicroft, Knapp, & Whiteford, 2007; World Health Organization, 2011) Because these types of measures may be freely available, short enough to lend themselves to easy translation and adaption, and do not

require extensively trained personnel to administer, self-report scales are often used to determine the burden of disorders and measure treatment impact in non-western, low-resource contexts.

2.7 Psychiatric assessment in global mental health

The most commonly used measures of psychopathology across cultures were developed in western clinical populations (Guillemin, Bombardier, & Beaton, 1993) and most were created using a foundation of classical test theory (CTT) (Table 1). Classical Test Theory fundamentally relies on the underlying population from which the items were derived. Research has shown that the use of CTT as a foundation for instrument development is heavily dependent on the population from which items were generated and scoring on a CTT created measure is specific to the individual test, thereby making comparisons across tests difficult (Hambleton, Waminathan, & Rogers, 1991). Moreover, the majority of the scales for assessing psychiatric illness were created in places with a history of psychology-related research and existing mental health systems already in place, such as the United States or Europe. Because of this history, most of the self-report scales commonly used in global mental health to measure depression include items that reflect the presentation of psychopathology in western clinical populations and DSM or ICD diagnostic criteria.

Table 2.1
Commonly used scales for depression in global mental health research with adults

Name of Schedule/Scale	Reference	Developed in what type of population	Examples of validation in LMIC
Composite International Diagnostic Interview (CIDI)	(Robins et al., 1988)	Psychiatric patients and general populations in United States and Europe	**Afghanistan** (Ventevogel et al. 2007); **Nepal** (Ghimire, Chardoul, Kessler, Axinn, & Adhikari, 2013); **Brazil** (Quintana, Gastal, Jorge, Miranda, & Andreoli, 2007)
Hopkins Symptom Checklist (HSCL)	(Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974)	Psychiatric outpatients in the United States	**Burmese refugees** (Haroz et al. 2014); **Cambodia** (Silove et al. 2007); **Northern Uganda** (Ertl et al. 2011); **Rwanda** (Bolton, 2001)
Center for Epidemiological Studies Depression (CES-D)	(Radloff, 1977)	General population in the United States	Armenians in Lebanon (Kazarian et al. 2009); India (Chokkanathan & Mohanty, 2013); Rwanda (Lacasse et al. 2014)
Patient Health Questionnaire (PHQ)	(Spitzer, Kroenke, & Williams, 1999)	Primary care patients in the United States	**Haiti** (Marc et al. 2014); Nigeria (Adewuya, Ola, & Afolabi, 2006); **Thailand** (Lotrakul, Sumrithe, & Saipanish, 2008);
Beck Depression Inventory (BDI)	(Beck, Ward, Mendelson, Mock, & Erbaugh, 1961)	Psychiatric outpatients receiving psychoanalytic psychotherapy in the United States	**Brazil** (Gomes-Oliveira, et al. 2012); **Turkey** (Kapci et al. 2008)
Hamilton Depression Rating Scale (HAM-D)	(Hamilton, 1960)	Psychiatric inpatients already diagnosed with an affective disorder	**China** (Zheng et al. 1988); **Turkey** (Akdemir et al. 2001)
Kessler's Psychological Distress Scale (K10)[a]	(Kessler et al., 2002)	Community samples in the United States; tested through government health surveys in the US and Australia	**Burkina Faso** (Baggaley et al. 2007); **Ethiopia** (Tesfeye et al. 2010); **Sri Lanka** (Wijeratne et al. 2011)

[a] Developed based on Item Response Theory

In global mental health research, comprehensive adaptation and testing of scales for assessing psychopathology is often the exception, rather than the norm. Many studies report simple translation and back translation of instruments with the expectation that these scales will preform the way they were intended across diverse settings. However, this is clearly not the case. Simple translation and back translation does not guarantee that items hold the same meaning in different languages or context, or that scales are valid and are measuring what they purport to measure (Bass, Bolton, & Murray, 2007; Bolton, 2001; Kohrt et al., 2011).

The items on the scales may not reflect presentation of true disorder in other contexts. Moreover, whether a scale is actually measuring a clinical severity level in need of treatment may vary from setting to setting. The use of previously established clinical cut-off scores that indicate disorder, may not signify the same level of distress across contexts (Bass et al., 2007). Instead, comprehensive adaptation of measures for psychopathology for use in culturally diverse settings should include a multi-stage process of translation, adaptation and testing in order to ensure accurate measurement of psychopathology (Bass et al., 2007; Kohrt et al., 2011).

Yet currently there is no single agreed upon process for comprehensive adaptation and validation of instruments (Kohrt et al., 2011). The only consensus that exists in the field, is that simple translation and back translation of instruments alone does not actually produce valid and reliable measures (Kohrt et al., 2011). Beyond that, the establishment of the validity of assessment instruments is still a challenge. Data generated by invalid instruments has serious implications for allocation of resources and could potentially result in inappropriate and harmful application of treatment (Kohrt et al., 2011; Wessells, 2009).

When considering the validity of screening instruments, several aspects of validity need to be considered. A major consideration is the relevance of the items to the target population. To address the issue of item relevance, some researchers have used qualitative research methods in scale development and adaptation in order to better incorporate locally relevant signs and symptoms of distress. The Applied Mental Health Research group (AMHR) at Johns Hopkins University follows a 5-step process in instrument development and adaptation. The first step involves conducting a brief ethnographic study aimed at identifying local signs, symptoms and syndromes related to mental distress. This step aids in the identification of appropriate established scales as candidates for adaptation (step 2). These scales are then translated using language that emerged during the qualitative study (step 3). In addition, if symptoms arise that are not included in the standard scales, then items are added to the instrument that are specifically relevant to the local context. Finally, the translated and adapted instrument are piloted (step 4) and

psychometrically tested (step 5) before wider use in a research study (Applied Mental Health Research Group, 2013).

While this approach provides a method for appropriately adapting existing scales and evaluating their validity in different contexts, it can be resource intensive (Hollifield, 2002). What has yet to be explored is the possibility that there may be near universal signs and symptoms of depression, which could be used to inform the creation of a depression measure that would have broad applicability across settings. This would be particularly important in situations where resources are scarce and there is not enough time to appropriately adapt instruments.

References

Altschule, M. D. (1965). Acedia: Its evolution from deadly sin to psychiatric syndrome. *The British Journal of Psychiatry, 111*(471), 117-119.

Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *Journal of Affective Disorders, 96*(1-2), 89-93.

Akdemir, A., Türkçapar, M., Örsel, S., Demirergi, N., Dag, I., & Özbay, M. (2001). Reliability and validity of the Turkish version of the hamilton depression rating scale. *Comprehensive Psychiatry, 42*(2), 161-165.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders fifth edition* (5th ed.*)*. Washington, DC: Author.

American Psychiatric Association. Task Force on Nomenclature and Statistics. (1980). *DSM III: Diagnostic and statistical manual of mental disorders*. Washington, DC: Author.

Anthony, J. C., Eaton, W. W., & Henderson, A. S. (1995). Looking to the future in psychiatric epidemiology. *Epidemiologic Reviews, 17*(1), 240-242.

Applied Mental Health Research Group (AMHR). (2013). *Design, implementation, monitoring and evaluation of cross-cultural trauma related mental health and psychosocial assistance programs: A user's manual for researchers and program implementers*. Retrieved from: http://www.jhsph.edu/research/centers-and-institutes/center-for-refugee-and-disaster-response/response_service/AMHR/dime

Baggaley, R., Ganaba, R., Filippi, V., Kere, M., Marshall, T., Sombie, I., . . . Patel, V. (2007). Short communication: Detecting depression after pregnancy: The validity of the K10 and K6 in Burkina Faso. *Tropical Medicine & International Health, 12*(10), 1225-1229.

Bass, J. K., Bolton, P. A., & Murray, L. K. (2007). Do not forget culture when studying mental health. *Lancet, 370*(9591), 918-918.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*(6), 561-571.

Bhugra, D., & Bhui, K. (2007). *Textbook of cultural psychiatry*. Cambridge, U.K: Cambridge University Press.

Bolton, P. (2001). Cross-cultural validity and reliability testing of a standard psychiatric assessment instrument without a gold standard. *The Journal of Nervous and Mental Disease, 189*(4), 238-242.

Bolton, P., Michalopoulos, L., Ahmed, A. M., Murray, L. K., & Bass, J. (2013). The mental health and psychosocial problems of survivors of torture and genocide in Kurdistan, northern Iraq: A brief qualitative study. *Torture : Quarterly Journal on Rehabilitation of Torture Victims and Prevention of Torture, 23*(1), 1-14.

Bolton, P., Neugebauer, R., & Ndogoni, L. (2002). Prevalence of depression in rural rwanda based on symptom and functional criteria. *The Journal of Nervous and Mental Disease, 190*(9), 631-637.

Bolton, P., Surkan, P. J., Gray, A. E., & Desmousseaux, M. (2012). The mental health and psychosocial effects of organized violence: A qualitative study in northern Haiti. *Transcultural Psychiatry, 49*(3-4), 590-612.

Bruckner, T. A., Scheffler, R. M., Shen, G., Yoon, J., Chisholm, D., Morris, J., . . . Saxena, S. (2011). The mental health workforce gap in low-and middle-income countries: A needs-based approach. *Bulletin of the World Health Organization, 89*(3), 184-194.

Carragher, N., Mewton, L., Slade, T., & Teesson, M. (2011). An item response analysis of the DSM-IV criteria for major depression: Findings from the Australian national survey of mental health and wellbeing. *Journal of Affective Disorders, 130*(1-2), 92-98.

Chokkanathan, S., & Mohanty, J. (2013). Factor structure of the CES-D scale among older adults in Chennai, India. *Aging & Mental Health, 17*(4), 517-525.

Choi, H., Fogg, L., Lee, E. E., & Wu, M. C. (2009). Evaluating differential item functioning of the CES-D scale according to caregiver status and cultural context in Korean women. *Journal of the American Psychiatric Nurses Association, 15*(4), 240-248.

Choi, Y., Mericle, A., & Karachi, T. W. (2006). Using rasch analysis to test the cross-cultural item equivalence of the harvard trauma questionnaire and the hopkins symptom checklist across Vietnamese and Cambodian immigrant mothers. *Journal of Applied Measurement, 7*(1), 16-38.

De Jong, J. T. V. M., & Van Ommeren, M. (2002). Toward a culture-informed epidemiology: Combining qualitative and quantitative research in transcultural contexts. *Transcultural Psychiatry, 39*(4), 422-433.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1974). The hopkins symptom checklist (HSCL): A self-report symptom inventory. *Behavioral Science, 19*(1), 1-15.

Draguns, J. G., & Tanaka-Matsumi, J. (2003). Assessment of psychopathology across and within cultures: Issues and findings. *Behaviour Research and Therapy, 41*(7), 755-776.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. New York: Russell Sage Foundation.

Ertl, V., Pfeiffer, A., Saile, R., Schauer, E., Elbert, T., & Neuner, F. (2010). Validation of a mental health assessment in an African conflict population. *Psychological Assessment, 22*(2), 318-324.

Ghalioungui, P. (1987). *The ebers papyrus: A New English Translation, Commentaries and Glossaries.* Academy of Scientific Research & Technology.

Ghimire, D. J., Chardoul, S., Kessler, R. C., Axinn, W. G., & Adhikari, B. P. (2013). Modifying and validating the composite international diagnostic interview (CIDI) for use in Nepal. *International Journal of Methods in Psychiatric Research, 22*(1), 71-81.

Gomes-Oliveira, M. H., Gorenstein, C., Lotufo Neto, F., Andrade, L. H., & Wang, Y. P. (2012). Validation of the brazilian portuguese version of the beck depression inventory-II in a community sample. *Revista Brasileira De Psiquiatria, 34*(4), 389-394.

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology, 46*(12), 1417-1432.

Hambleton, R. K., Waminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (1st ed.). California: Sage Publications Inc.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23*(1), 56-62.

Hollifield, M. (2002). Accurate measure in cultural psychiatry: Will we pay the costs? *Transcultural Psychiatry, 39*, 419–421.

Hwu, H., & Compton, W. M. (1994). Comparison of major epidemiological surveys using the diagnostic interview schedule. *International Review of Psychiatry, 6*(4), 309-327.

Jackson, S. W. (1969). Galen—on mental disorders. *Journal of the History of the Behavioral Sciences, 5*(4), 365-384.

Jackson, S. W. (1985). Acedia the sin and its relationship to sorrow and melancholia. In A. Kleinman & B. Good (Eds.) *Culture and Depression: Studies in the Anthropology and Cross-Cultural Psychiatry of Affect and Disorder* (43-62). California: University of California Press.

Jilek, W. G. (1995). Emil kraepelin and comparative sociocultural psychiatry. *European Archives of Psychiatry and Clinical Neuroscience, 245*(4), 231-238.

Jones, W. H. S., Withington, E. T., & Potter, P. (1928). *Hippocrates*. Loeb Classical Library. Cambridge MA: Harvard University Press.

Kaiser, B. N., McLean, K. E., Kohrt, B. A., Hagaman, A. K., Wagenaar, B. H., Khoury, N. M., & Keys, H. M. (2014). Reflechi twop--thinking too much: Description of a cultural syndrome in Haiti's central plateau. *Culture, Medicine and Psychiatry, 38*(3), 448-472.

Kapci, E. G., Uslu, R., Turkcapar, H., & Karaoglan, A. (2008). Beck depression inventory II: Evaluation of the psychometric properties and cut-off points in a Turkish adult population. *Depression and Anxiety, 25*(10), 104-110.

Kazarian, S. S. (2009). Validation of the armenian center for epidemiological studies depression scale (CES-D) among ethnic Armenians in Lebanon. *The International Journal of Social Psychiatry, 55*(5), 442-448.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S., . . . Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine, 32*(06), 959-976.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., . . . Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: Results from the national comorbidity survey. *Archives of General Psychiatry, 51*(1), 8-19.

Kirmayer, L. J. (1984). Culture, affect and somatization part I. *Transcultural Psychiatry, 21*(3), 159-188.

Kirmayer, L. J. (2007). Cultural psychiatry in historical perspective. In D. Bhuga & K. Bhui (Eds.) *Textbook of Cultural* Psychiatry (3-19). Cambridge, U.K.: Cambridge University Press.

Kleinman, A. (1982). Neurasthenia and depression: A study of somatization and culture in China. *Culture, Medicine and Psychiatry, 6*(2), 117-190.

Kohrt, B. A., & Hruschka, D. J. (2010). Nepali concepts of psychological trauma: The role of idioms of distress, ethnopsychology and ethnophysiology in alleviating suffering and preventing stigma. *Culture, Medicine, and Psychiatry, 34*(2), 322-352.

Kohrt, B. A., Jordans, M. J. D., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). Validation of cross-cultural child mental health and psychosocial research instruments: Adapting the depression self-rating scale and child PTSD symptom scale in nepal. *BMC Psychiatry, 11-27.*

Kohrt, B. A., Rasmussen, A., Kaiser, B. N., Haroz, E. E., Maharjan, S. M., Mutamba, B. B., . . . Hinton, D. E. (2014). Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations for global mental health epidemiology. *International Journal of Epidemiology, 43*(2), 365-406.

Kraepelin, E. (1904). Vergleichende psychiatrie. *Zentralbl Nervenheilkd Psychiatry, 27*, 433–437.

Lacasse, J. J., Forgeard, M. J., Jayawickreme, N., & Jayawickreme, E. (2014). The factor structure of the CES-D in a sample of Rwandan genocide survivors. *Social Psychiatry and Psychiatric Epidemiology, 49*(3), 459-465.

Lotrakul, M., Sumrithe, S., & Saipanish, R. (2008). Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry, 8*, 46-53.

Maes, K. C., Kohrt, B. A., & Closser, S. (2010). Culture, status and context in community health worker pay: Pitfalls and opportunities for policy research. A commentary on Glenton et al. (2010). *Social Science & Medicine, 71*(8), 1375-1380.

Manson, S. M., Shore, J. H., & Bloom, J. D. (1985). The depressive experience in American Indian communities: A challenge for psychiatric theory and diagnosis. In A. Kleinman & B.

Good (Eds.) *Culture and Depression: Studies in the Anthropology and Cross-Cultural Psychiatry of Affect and Disorder* (331-368). California: University of California Press.

Marc, L. G., Henderson, W. R., Desrosiers, A., Testa, M. A., Jean, S. E., & Akom, E. E. (2014). Reliability and validity of the Haitian creole PHQ-9. *Journal of General Internal Medicine, 29*(12), 1679-1686.

Marsella, A. J., Friedman, M. J., Gerrity, E. T., & Scurfield, R. M. (1996). Ethnocultural aspects of PTSD: Some closing thoughts (529-538). Washington, D.C.: American Psychological Association.

Miller, K. E., Omidian, P., Quraishy, A. S., Quraishy, N., Nasiry, M. N., Nasiry, S., . . . Yaqubi, A. A. (2006). The Afghan symptom checklist: A culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry, 76*(4), 423-433.

Murphy, H. B. (1982). Comparative psychiatry: The international and intercultural distribution of mental illness. *Monographien Aus Dem Gesamtgebiete Der Psychiatrie, 28,* 1-327.

Murphy, J. M. (1976). Psychiatric labeling in cross-cultural perspective. *Science, 191*(4231), 1019-1028.

Murphy, J. M., Tsuang, M., & Tohen, M. (2002). Symptom scales and diagnostic schedules in adult psychiatry. In Tsuang, M.T., Tohen, M. & Jones, P.B. (Eds) *Textbook in Psychiatric Epidemiology,* (273-332). New York, NY: John Wiley & Sons, Ltd.

Nichter, M. (2010). Idioms of distress revisited. *Culture, Medicine, and Psychiatry, 34*(2), 401-416.

Olino, T. M., Yu, L., Klein, D. N., Rohde, P., Seeley, J. R., Pilkonis, P. A., & Lewinsohn, P. M. (2012). Measuring depression using item response theory: An examination of three measures of depressive symptomatology. *International Journal of Methods in Psychiatric Research, 21*(1), 76-85.

Patel, V., Simunyu, E., Gwanzura, F., Lewis, G., & Mann, A. (1997). The shona symptom questionnaire: The development of an indigenous measure of common mental disorders in harare. *Acta Psychiatrica Scandinavica, 95*(6), 469-475.

Patel, V., & Prince, M. (2010). Global mental health: A new global health field comes of age. *JAMA, 303*(19), 1976-1977.

Quintana, M. I., Gastal, F. L., Jorge, M. R., Miranda, C. T., & Andreoli, S. B. (2007). Validity and limitations of the Brazilian version of the composite international diagnostic interview (CIDI 2.1). *Revista Brasileira De Psiquiatria, 29*(1), 18-22.

Radloff, L. S. (1977). The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Rasmussen, A., Katoni, B., Keller, A. S., & Wilkinson, J. (2011). Posttraumatic idioms of distress among Darfur refugees: Hozun and majnun. *Transcultural Psychiatry, 48*(4), 392-415.

Regier, D. A., Myers, J. K., Kramer, M., Robins, L. N., Blazer, D. G., Hough, R. L., . . . Locke, B. Z. (1984). The NIMH epidemiologic catchment area program: Historical context, major objectives, and study population characteristics. *Archives of General Psychiatry, 41*(10), 934-941.

Robert F. DeVellis. (2012). *Scale development: Theory and applications* (3rd ed.). California: SAGE publications, Inc.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National institute of mental health diagnostic interview schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38*(4), 381-389.

Robins, L. N., Wing, J., Wittchen, H. U., Helzer, J. E., Babor, T. F., Burke, J., . . . Regier, D. A. (1988). The composite international diagnostic interview: An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry, 45*(12), 1069-1077.

Sartorius, N. (1983). Depressive disorders in different cultures: Report on the WHO collaborative study on standardized assessment of depressive disorders. Geneva: World Health Organization.

Saxena, S., Thornicroft, G., Knapp, M., & Whiteford, H. (2007). Resources for mental health: Scarcity, inequity, and inefficiency. *The Lancet, 370*(9590), 878-889.

Silove, D., Manicavasagar, V., Mollica, R., Thai, M., Khiek, D., Lavelle, J., & Tor, S. (2007). Screening for depression and PTSD in a Cambodian population unaffected by war. *The Journal of Nervous and Mental Disease, 195*, 152-157.

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD. *JAMA, 282*(18), 1737-1744.

Summerfield, D. (2000). War and mental health: A brief overview. *BMJ (Clinical Research Ed.), 321*(7255), 232-235.

Tesfaye, M., Hanlon, C., Wondimagegn, D., & Alem, A. (2010). Detecting postnatal common mental disorders in addis ababa, ethiopia: Validation of the edinburgh postnatal depression scale and kessler scales. *Journal of Affective Disorders, 122*(1), 102-108.

Ventevogel, P., De Vries, G., Scholte, W. F., Shinwari, N. R., Faiz, H., Nassery, R., ... & Olff, M. (2007). Properties of the Hopkins Symptom Checklist-25 (HSCL-25) and the Self-Reporting Questionnaire (SRQ-20) as screening instruments used in primary care in Afghanistan. *Social Psychiatry and Psychiatric Epidemiology*, *42*(4), 328-335.

Wessells, M. G. (2009). Do no harm: Toward contextually appropriate psychosocial support in international emergencies. *The American Psychologist, 64*(8), 842-854.

Wijeratne, L., Williams, S., Rodrigo, M., Peris, M., Kawamura, N., & Wickremasinghe, A. (2011). Validation of the Kessler's psychological distress scale among the Sinhalese population in Sri Lanka. *South Asian Journal of Psychiatry,* 21-25.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

World Health Organization. (1973). Report of the international pilot study of schizophrenia. Geneva: World Health Organization.

World Health Organization. (2011). *Mental health atlas, 2011.* Geneva: World Health Organization.

Zheng, Y. P., Zhao, J. P., Phillips, M., Liu, J. B., Cai, M. F., Sun, S. Q., & Huang, M. F. (1988). Validity and reliability of the Chinese hamilton depression rating scale. *The British Journal of Psychiatry, 152*, 660-664.

Chapter 3. Methods

This chapter aims to describe the methodological considerations taken when conducting the analyses of all three aims of this dissertation. Specifically it focuses on issues related to systematic reviews of qualitative data (Aim 1); explanation and interpretation of Item Response Theory (IRT) analyses (Aim 2); and methods used for reliability and validity testing of self-report instruments to measure psychopathology (Aim 3).

3.1 Aim 1 Methods

Background on systematic reviews of qualitative data

According to the Cochrane Library a systematic review "attempts to identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a given research question," (Higgens & Green, 2011). Systematic review methodology has largely been developed for the review of intervention trials, while the methodology for systematically reviewing non-experimental research, such as observational or qualitative studies, is still emerging (Dixon-Woods, Fitzpatrick, & Roberts, 2008; Harden et al., 2004). To date there exists no agreed upon methodology for the review of qualitative studies. Yet, this type of review is important, as evidence generated from these types of studies could be used to explain, confirm and/or refute results from other studies (Thomas & Harden, 2008).

Reviews of qualitative data are different from traditional systematic reviews. Qualitative research is specific to the population, time, and the context under study. By design, this type of research often lacks generalizability. Thus, the decontextualizing and synthesizing of findings across studies, as one would do in a systematic review, appears to go against the very nature of qualitative research. However, in many qualitative research studies the argument that the findings from such studies have the potential to influence policy and practice has also been made. In order to balance these two viewpoints—lack of generalizability and influence on policy and practice—

review strategies of qualitative studies that acknowledge the complexity and contextualization of the research are necessary (Thomas & Harden, 2008).

Challenges with systematic reviews of qualitative data

One of the first major challenges in reviews of qualitative data is identifying articles to include. In traditional systematic reviews, the aim is to locate all relevant studies so as to maximize the inferences gathered from statistical analysis. However, in reviews of qualitative data, because statistical inferences are not paramount, authors have taken many approaches to finding and locating the necessary articles. The most common approaches involve either literature searches aimed at maximizing the likelihood of identifying relevant articles, while excluding articles that are irrelevant to the research question (Shaw et al., 2004); or aiming for conceptual saturation, maximizing variability, and finding studies that act as negative cases (Thomas & Harden, 2008).

Over identification of articles can be a real problem. For example, Shaw and colleagues (Shaw et al., 2004) combined Cochrane review search strategies with utilization of alternative methods to generate search terms, such as thesaurus terms, free-text, and broad-based terms to identify qualitative research related to breast-feeding. These strategies resulted in a large number of "false-positives" or texts that were not relevant to the research questions (Shaw et al., 2004). Harden et al. (2004) conducted a systematic reviews to investigate the barriers and facilitators of mental health, physical activity and healthy eating among young people in England. The authors designed a strategy that began with identifying which types of studies had the potential to answer the research question. After generating search terms, the authors identified 18686, potential articles to review. These were subsequently reduced to 510 studies, through a mapping exercise. Further reduction was done through meetings with key-stakeholders to identify a subset of these studies for in-depth review.

Evaluating the rigor of qualitative research is also a challenge. In general, there is no consensus in the field about how to assess the quality of qualitative studies. Some researchers have specified certain criteria that focused on the context, sample, data collection techniques, background of the researcher, and importance (Cobb & Hagemaster, 1987; Mays & Pope, 1995; Thomas & Harden, 2008). Others have developed criteria for reporting of qualitative data that could be used to evaluate the quality of the studies as well (Tong, Sainsbury, & Craig, 2007). However, because of the lack of data supporting one approach over the other, and the potential consequence of excluding studies that may add relevant information, a consensus strategy of what constitutes rigorous qualitative methodology has not been agreed upon (Thomas & Harden, 2008).

Both over identification and evaluation of rigor, were concerns for Aim 1's systematic review of symptoms of depression mentioned in qualitative research. However, while over identification can be burdensome on the researcher, it can also be necessary, at least initially, in order to ensure that nothing important is missed. Thus, for Aim 1, the goal was to identify as many articles as possible that potentially met the inclusion criteria. Second, because of the lack of consensus on what constitutes a rigorous qualitative study, evaluation of the methods used in each study included in the review, was not done. However, when possible, the methods used were noted, but ultimately not evaluated.

3.1.a Analysis for Aim 1

The review systematically examined qualitative research related to depression to identify a set of signs and symptoms on depressive-like syndromes in a range of cultures and settings (Aim 1). The literature review followed PRISMA guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009). Nine online databases were searched in order to identify all articles that mentioned a sign or symptom related to depression. Databases included Pubmed, Web of Science,

PsychInfo, Scopus, Embase, Anthrosource, Anthropology Plus, Global Health and Sociological Abstracts. Additionally, Google Scholar was used after the initial search to determine whether particularly relevant articles were missed during the initial search, and any article missed was added to the results. The search was conducted using a two-staged approach. The first search used the following terms: "depression," "depressive disorder," "melancholia," and "depressive disorder, major." The resulting set of articles was then searched (2nd stage of search) for the following terms: "anthropology," "qualitative," "ethnography," "cross-cultural comparison," "ethnopsychology," cultural characteristics," "cross cultur*," "phenomenology," and "idioms of distress." For both stages, study titles, abstracts, and subjects were search and MeSH terms were used when possible. If a review article was identified during the search, the bibliography of the review article was also searched for possible relevant citations, and if any articles not already identified were listed in the bibliography, these were included for full-text review. In addition, programmatic reports that could be located but had not been published in the peer-literature, were added.

Inclusion criteria for each article was as follows: 1) represent qualitative research; 2) have depression as the main focus of the research; 3) include information on symptoms of depression; 4) be written in English; and 5) report on a study population of adults between the ages of 18-65. Any article that reported data on only one individual or on a small series of case-studies was omitted (exclusion criteria). Due to the lack of agreement on proper appraisal of the rigor of qualitative research (Thomas & Magilvy, 2011; Tobin & Begley, 2004) and the desire to be as overly inclusive as possible, articles were not appraised on the quality of the original research.

For each article that met inclusion/exclusion criteria, symptoms associated with depression that were mentioned in the text were extracted. Other information extracted from the final set of eligible articles included: 1) sex of the study population; 2) region of the world; 3) nationality and/or ethnicity; 4) religious distinction if available; 5) class distinction if available; 6)

whether the study took place in the perinatal context, or the context of war, trauma or displacement; 7) whether the study took place in an urban or rural location if available; and 8) which qualitative research methods were used.

Extracted symptoms of depression were coded using a priori and emergent coding. A priori codes were based on the symptoms of Major Depressive Disorder included in the DSM-V (American Psychiatric Association, 2013) (Table 1). Emergent coding involved reviewing those symptoms that did not reflect DSM-V diagnostic criteria and grouping together symptoms representing the same idea. After all articles had been reviewed and respective symptoms coded, a quantitative dataset was compiled. The dataset included rows for each study population and columns with the name of symptom code. As some articles reported on multiple study populations, the number of rows in the dataset was greater than the number of articles included in the review. For each study population (row), whether or not the symptom was reported was marked as present or not present (dichotomous). This quantitative dataset was then analyzed to examine the frequencies of symptoms overall, by region, by gender, and by contextual variable (perinatal context or context of war/trauma/displacement).

Table 3.1

A-prior codes based on DSM-5 diagnostic criteria for Major Depressive Disorder

1. Depressed mood
2. Diminished interest or pleasure
3. Significant weight loss or weight gain
4. Insomnia or hypersomnia
5. Psychomotor agitation or slowing
6. Fatigue or loss of energy
7. Worthlessness or inappropriate guilt

8. Diminished ability to think or concentrate, or indecisiveness

9. Recurrent thoughts of death/suicidal ideation

10. Functional impairment

11. Irritability

3.2 Aim 2 Methods

Background on Classical Test Theory vs. Item Response Theory

Classical Test Theory (CTT) has formed the basis for psychological scale development and measurement for over 100 years (DeVellis, 2012). CTT posits that an individual's observed score is determined by their true score plus measurement error (Figure 1). A true score is defined as the average score on a test if the individual took the test an infinite number of times. The assumptions of CTT are as follows: 1) error associated with individual items is randomly distributed with a mean of zero when aggregated across large populations; 2) item errors are independent of error on other items; and 3) error terms are not correlated with the true score of the latent variable.

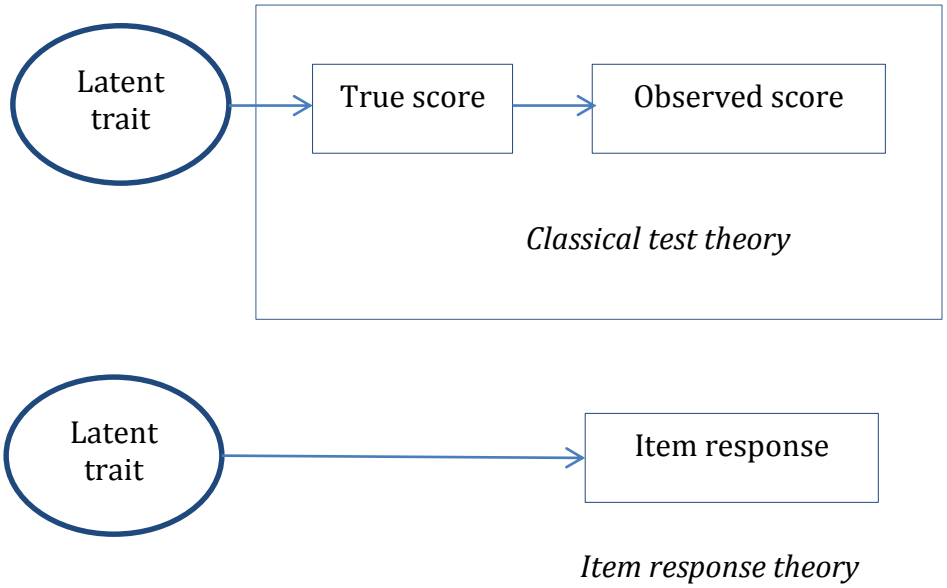
CTT is limited in its utility in areas of scale development and refinement, particularly when applying scales in different settings. In CTT the characteristics of the scales and items are heavily dependent on the population from which they were derived, making them of limited use in dissimilar populations and/or requiring extensive retesting in different populations to establish validity (Hambleton et al., 1991).

An alternative approach to scale development and refinement is the use of Item Response Theory (IRT). IRT is a type of latent variable analysis and is referred to as a latent trait model. Latent trait models are used in situations where the latent variable (i.e. unobserved variable) is thought to be continuous, but observed indicators are categorical. In IRT an individual's response to an item is directly predicted by their latent traits or abilities (Figure 1). IRT was initially

developed for use in educational testing, but can be applied to psychometric testing situations as well.

Figure 3.1

Path diagrams for classical test theory vs. item response theory



IRT utilizes a conditional probability framework such that, as the level of the latent trait increases the probability of endorsing the item also increases. More specifically, by conditioning on an individual's latent trait (ɵ), characteristics of test items can be described independently from the sample to whom it was administered. Comparisons of CTT and IRT can be seen in Table 2 (Hambleton et al., 1991).

Table 3.2

Comparison of CTT and IRT

	Classical Test Theory	Item Response Theory
Basic equation	O = T + error	$P_i(\theta) = \frac{e^{Da_{i(\theta - b_i)}}}{1+e^{Da_{i(\theta - b_i)}}}$
Reliability	Assumed constant across ɵ	Conditional on ɵ
Scoring	Test dependent	Not test dependent
Item properties	Sample dependent	Not sample dependent

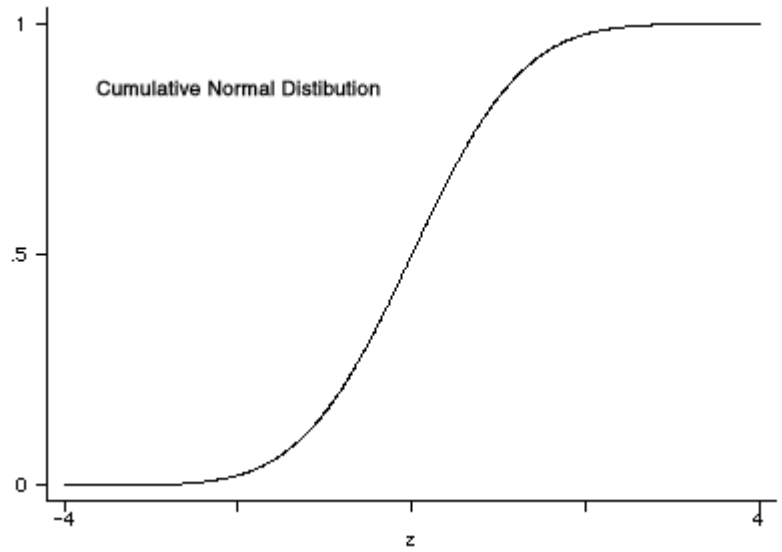
* D is a scaling factor used to make the logistic function as close as possible to a cumulative normal as possible

**ɵ = latent trait

In IRT the relationship between a person's latent trait (ɵ) and the item response is explained by an Item Characteristic Curve (ICC), which is a cumulative probability function that depicts the probability of endorsing a certain item increasing as the level of the latent trait increases. This function is assumed to follow a cumulative normal distribution (Figure 2).

Figure 3.2

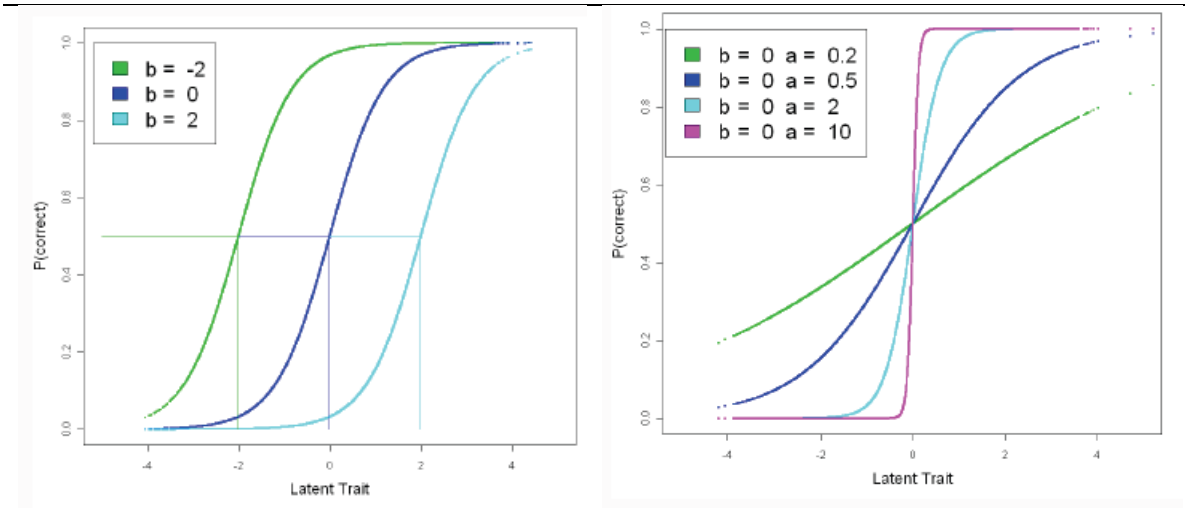
Cumulative Normal Distribution



ICCs can have up to 3 parameters depending on the type of model; 1) item difficulty parameter (or location parameter represented by the notation *b*); 2) item discrimination parameter (*a*); and 3) a pseudo-chance level parameter (*c*). The item difficulty parameter (*b*) represents the point on the latent trait scale where the probability of endorsing an item is 0.50. For example, an item that has a *b* parameter equal to -1 might be an item that measures less severe depression, then an item that has a *b* parameter equal to 1, which would be an item that measures more severe depression. The item discrimination parameter (*a*) describes the ability of an item to discriminate between lower and higher levels of the latent trait value, and is equivalent to a factor loading. For example, an item with high discrimination (*a*) can better distinguish between individuals at different levels of the trait in the region of the item location.. The pseudo-chance level parameter (*c*) is normally only used in testing situations when chance can play a role in performance such as in standardized tests. The *c* parameter represents the probability of endorsing items for respondents with very low levels of the latent trait (Figure 3) (Hambleton et al., 1991).

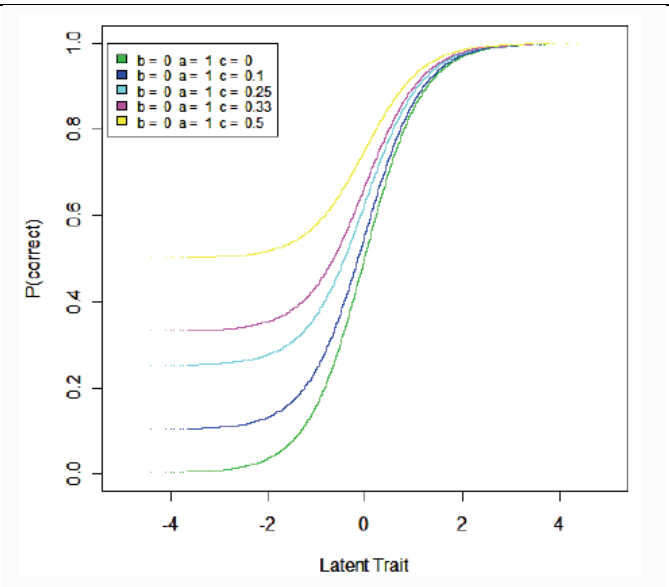
Figure 3.3

Item characteristic curves for 1, 2 and 3 parameter models



1 parameter model: Item difficulty (*b*)

2 parameter model: Item discrimination (*a*) & Item difficulty (*b*)



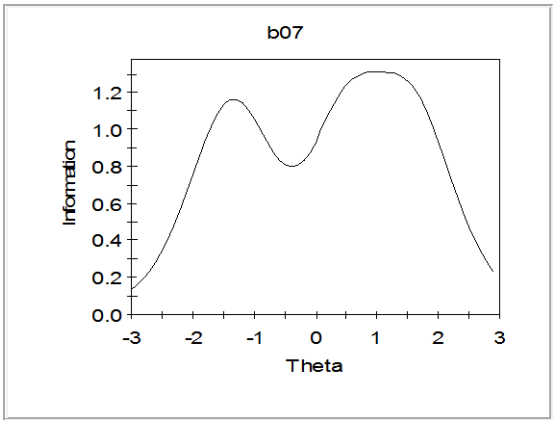
3 parameter model: Pseudo-chance level parameter (*c*), Item discrimination (*a*) & Item difficulty (*b*)

IRT analyses also generate item information curves and test information functions (TIF) that can help in evaluating item level performance (Figure 4). Item information, is similar to reliability in CTT, and indicates how precisely the item estimates the level of the latent trait. However, in contrast to reliability in CTT, item information can vary over the range of the latent trait. For example, an item related to suicide ideation, may be most reliable for individuals with high levels of depression, but not reliable for individuals with lower levels of depression. The item information curve depicts this varying reliability. Test information indicates how well the test does in estimating the latent trait (ɵ) over the entire range of latent trait scores. Test Information Functions (TIFs), are generated by summing the item information parameters and are influenced both by the number and quality of items. Flat TIFs, suggest the test has comparable precision over the range of the latent trait, while peaked TIFs suggest unequal precision. For TIFs, a peak that indicates information greater than 20 is considered excellent and corresponds to a reliability of 0.95 (Hambleton et al., 1991).

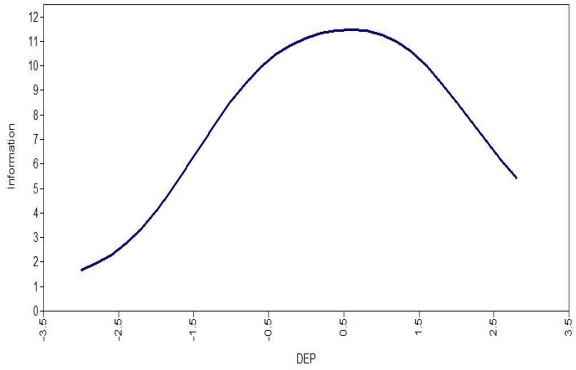
Figure 3.4

Item information and test information curves

Item Information Curve



Test Information Curve



IRT models include several assumptions about the data under analysis. IRT models are predicated on the idea that the probability of endorsing an item in a certain way depends on both the person's latent trait and characteristics of the items (assumption 1). This contrasts with the assumption of CTT, which assumes that a person's response is only dependent on the characteristics of the item itself. The second assumption relates to local independence (or conditional independence), which states that after conditioning on the value of the latent trait, there is no association between the responses to items (Hambleton et al., 1991).

The third assumption involves, the assumption of unidimensionality, meaning that only one latent trait is being measured. Formally, unidimensional IRT models are models for which a dominant latent trait adequately explains a person's performance on a questionnaire. In practice, unidimensionality can be hard to achieve as many factors, including personality, cognitive abilities, and test-taking environment, often influence the way a person responds to a test/questionnaire. However, to meet the assumption of unidimensionality, it is only required that one factor/latent trait be dominant in influencing test/questionnaire performance (Hambleton et al., 1991).

In recent years, multidimensional IRT models have been developed. In these models, more than one latent trait is thought to explain responses on a test/questionnaire. Multidimensional models expressly model the different factors/components that underlie a person's response to items (Hambleton et al., 1991). Multidimensional models may be particularly important in psychometrics, as many latent traits have multiple dimensions that explain item responses.

Studies have examined the effects of using a unidimensional IRT model in modeling item response data that is not strictly unidimensional (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Kirisci, Hsu, & Yu, 2001). These studies have found that if there is one predominant general factor and other much smaller factors, the use of multidimensional models does not significantly change the parameter estimates. However, if there is strong evidence of multiple factors, then parameter estimates based on a unidimensional model do not accurately represent the true data, but are rather more representative of the dominant factor (Gibbons, Immekus, Bock, & Gibbons, 2007).

In the IRT literature, several methods have been suggested for checking the dimensionality of the underlying latent trait. These methods include examination of factor eigenvalues, investigation of item local independence within different intervals on the latent trait scale, and fitting a nonlinear one-factor analysis model and examining the residuals (Hambleton et al., 1991). Building off the first method related to eigenvalues, the use of a Principle Components Analysis (PCA) with parallel analysis, can be used to investigate the underlying number of factors and thus the dimensionality of the data.

Differential Item Functioning

Differential item functioning (DIF), often referred to as item bias, is an important consideration in scale development. An item shows DIF if individuals with the same value of the

latent trait have different probabilities of endorsing the item because of other characteristics of the individual. Concern about DIF, originally arose in standardized testing situations, in which African American respondents systematically answered certain items differently than White Americans. In the context of psychometric scales DIF could be affected by gender, age, or another variable that influences the responses to items despite similar latent trait values.

DIF is important in cross-cultural research as well. Despite general consensus that depression is a universal disorder, the symptoms of depression may differ across cultures and contexts. Thus individuals from different cultures or contexts who have the same degree of underlying depressive illness may respond to an item in a depression scale in systematically different ways. Evaluation of DIF within the IRT framework allows for detecting the influence of various variables on response to items aimed at measuring depression (Hambleton et al., 1991).

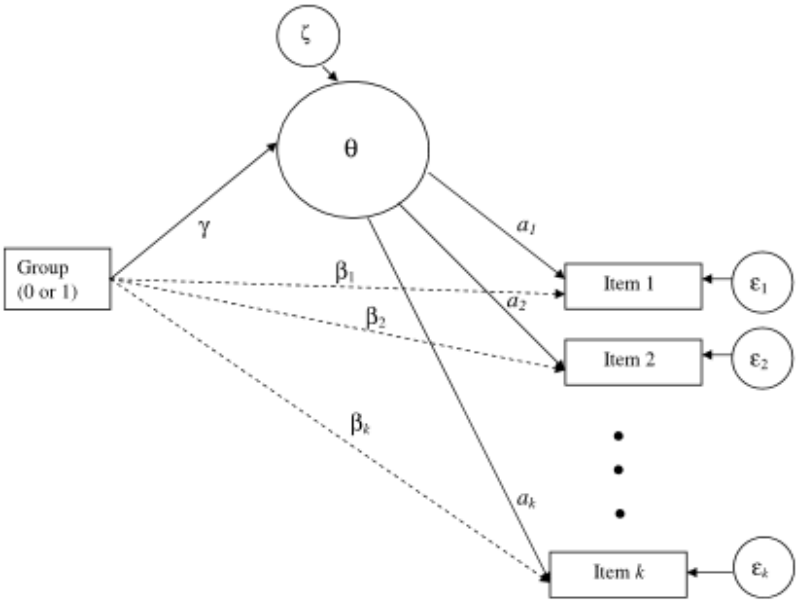
There are two types of DIF that can be present for an item: non-uniform DIF and uniform DIF. Non-uniform DIF is DIF in the item discrimination parameters (a) and represents an interaction between the trait level, group membership and the item response (Crane, Belle, & Larson, 2004). Uniform DIF, or DIF in the location parameters, can be thought of as confounding, and is present when differences in responses to items can be found at all levels of the latent trait (Crane et al., 2004).

Investigation into DIF can be done in a number of different ways: non-parametric methods such as Mantel-Haenszel and standardization techniques; or parametric methods such as ordinal and logistical regression (Teresi, 2006). Multiple Indicators Multiple Causes models (MIMIC; Gallo, Anthony, & Muthen, 1994; B. Muthén, 1985), is a parametric method that can be used to identify both non-uniform and uniform DIF (Woods, 2009; Woods & Grimm, 2011). MIMIC models are a type of structural equation model (SEM). MIMIC models model the direct effect of the group variable (in this case country membership) on the latent variable (e.g. depression) and each measured item simultaneously (Figure 5). The measure of DIF is then the coefficient of the path relating the group variable to the item, after controlling for the impact of

the group variable on the latent variable and the latent variable on the measured item. To detect DIF, MIMIC models use an iterative process, whereby the direct effect of the group variable on the first item is added to the model and modification indices are examined (which indicate whether there is a significant relationship between group membership and the measured item). A second model is then run, now adding the direct effect of the group variable on the first and second items and again examining modification indices. This process is repeated for every item in the scale and the items with significant modification indices after all iterations, are identified as showing DIF, while those with non-significant modification indices are considered to be free of DIF.

Figure 3.5

Path diagram for MIMIC model



3.2.a Analysis for Aim 2

For Aim 2, IRT analysis was performed to investigate item level parameters for a commonly used depression screening measure, the Hopkins Symptom Checklist 15-item version

depression subscale (HSCL-15; (Hesbacher, 1980; Winokur, Winokur, Rickels, & Cox, 1984), using data collected from eight different study settings. For each item on the HSCL respondents indicated how often they have felt a certain way over a certain recall period (recall period ranged from 1 to 4 weeks across studies) utilizing a ordinal response scale. Response options ranged from *0 – 3* with 0 = "None of the time" and 3 = "Almost all the time."

Sample

The sample involved a combination of data from eight study settings. All of the samples were from research studies conducted by the Applied Mental Health Research Group (AMHR) at Johns Hopkins Bloomberg School of Mental Health in conjunction with local partners at each of the sites. These studies shared similar research protocols utilizing the DIME process (Design, Implementation, Monitoring and Evaluation). Briefly, the DIME approach involves an initial qualitative study to understand relevant problems faced by a specific study population. This qualitative research goes on to inform instrument development and selection of an appropriate intervention to help address the identified mental health problems. Mental health instruments developed as part of the DIME process are tested and validated in each context and then used as screening measures for determining intervention eligibility and intervention impact (Applied Mental Health Research Group, 2013). Information about each of the datasets is included in Table 3.

Table 3.3

Description of data included in IRT analysis

Study Setting	***Type of study***	***N***
Colombia*	Screening, Validity	1263
Northern Iraq/Kurdistan		
Dohuk	Clinical Monitoring	294
Erbil and Sulaimaniya	Clinical Monitoring	680
Indonesia*	Screening, Validity	588
Southern Iraq	Validity	149
Rwanda	Epidemiologic study	368
Thailand*	Screening, Validity	803
Uganda	Epidemiologic study	587
TOTAL		4732

*Observations missing all item-level data for depression were dropped from this table and all subsequent analyses (Colombia: $n = 1$; Indonesia: $n = 1$; Thailand: $n = 15$)

All data were collected among trauma-affected populations. The samples from Colombia, Indonesia and Thailand, were collected as part of validation studies to test the reliability and validity of an adapted instruments and as screening samples for a Randomized Control Trial (RCT) of psychotherapeutic interventions. The Colombia study included participants, who were recruited in partnership with the CISALVA Institute at *Universidad del Valle* in Colombia. Participants were Afro-Colombian in origin and had been internally displaced due to armed conflict. The Indonesia study involved torture affected adults in Ache, Indonesia (Bass et al., 2012). The Thailand study included Burmese participants who were living on the Thai/Burma border and had experienced torture or trauma (Bolton et al., 2014a; Haroz et al., 2014).

The samples from Kurdistan, Dohuk and Erbil/Sulaimaniya, come from a clinic-based monitoring system, which was established as part of the implementation of an RCT of psychotherapeutic intervention to reduce the severity of mental health symptoms experienced by torture survivors in Kurdistan, Iraq. The data include participants who were eligible for the RCT as well as all clients who were assessed at the clinic but not found to be eligible for the trial (Bolton et al., 2014b).

The data from Rwanda and Uganda came from population-based surveys of trauma-affected adults in each of these countries (Bolton, Wilk, & Ndogoni, 2004; Bolton, Neugebauer, & Ndogoni, 2002). Finally, the data from Southern Iraq were from a validation study of an instrument designed to measure psychological distress among victims of torture (Weiss & Bolton, 2010).

Data were collected as part of several types of studies. *Validity studies* involved testing of instruments to measure mental health problems to establish the instruments reliability and validity in the local context. *Screening studies* took place in the context of screening for randomized control trials (RCT) of psychotherapeutic interventions. Screening involved determining whether an individual met a level of symptom severity (established during previous validity studies) to likely benefit from services. The data from these studies represent a mix of people who did and did not meet severity criteria for the RCTs. *Clinical monitoring studies* involved screening of persons that presented to clinic mental health staff. For the Kurdistan studies, the same instrument was used to determine eligibility for the RCTs as well as used for regular clinical intakes for all adults who sought services. *Epidemiologic studies* involved representative samples and were aimed at estimating prevalence of selected mental health disorders.

IRT analysis

To account for the ordered response categories on the HSCL, A Samejima graded response model (Samejima, 1997) was fit to the data. Item location and discrimination

parameters, as well as item information curves were examined across all countries and within each country separately. In addition, all items on the HSCL-15 were evaluated for uniform and non-uniform DIF by country using MIMIC models. The impact of item-level DIF on aggregate estimates of depression, was examined by comparing estimates of depression that accounted for item-level DIF, to estimates of depression that did not account for item level DIF. All statistical analysis will be done using STATA, 11 (StataCorp, 2009), Mplus version 7 (L. Muthén & Muthén, 2012), and IRTPRO (Cai, Du Toit, & Thissen, 2011).

3.3 Aim 3 Methods

Aim 3 involved testing the reliability and validity of a depression measure created based on findings from Aim 1 and Aim 2 in a sample of approximately $N = 150$ individuals recruited through community healthcare clinics in Yangon, Myanmar. To be included in the study, participants had to be a patient at either of the clinics, be literate in Burmese, and be over the age of 18. Participants were excluded if they showed signs of active psychosis or presence of a major developmental delay.

The depression scale created based on evidence from aims 1 and 2, is titled the International Depression Symptom Scale (IDSS) and consists of a total of 29 items, 27 of which are used for scoring purposes. Two items are not intended to be included in the summary score as these items while clinically important, were not supported for inclusion by Aims 1 and 2. These items related to suicide ideation and impaired functioning. The IDSS was designed to be flexible enough to incorporate locally relevant signs and symptoms. Thus, the IDSS consists of a global measure (IDSS-G), which includes the core set of 29 items; and a local measure (IDSS-L) which includes the 29 core items as well as any additional items reflecting locally specific symptoms of depression. For the purpose of Aim 3, just the global scale was tested, although a local scale was created with the addition of one item (“disappointment”) and results for the IDSS-L can be found in Appendix F.

Respondents were asked how often in the last two weeks had they experienced each of the symptoms on the IDSS. Response options ranged from 0 “none of the time” to 3 “almost all the time.” Average scores for the IDSS were generated by summing across the 27 items and dividing the sum by 27. In addition to the IDSS, participants completed the Patient Health Questionnaire-9 (PHQ-9; Kroenke & Spitzer, 2002) and a measure of functional impairment which had been developed in a similar population (Haroz et al., 2014).

The reliability, validity and clinical utility of the IDSS were evaluated. Reliability was assessed using internal consistency reliability Cronbach’s Alpha (α) (Cronbach, 1951) and item analysis; principal components analysis (PCA) with parallel analysis and exploratory factor analysis (EFA); and evaluation of test-retest reliability and inter-rater reliability. Validity was assessed through evaluation of face validity, external construct validity, criterion validity, and incremental validity. In addition, clinical utility was evaluated using Receiver Operating Curves (ROC) to look at area under the curve (AUC) and determine optimal cutoff points on the IDSS to maximize sensitivity and specificity of the measure to identify individuals with either a mix of depression (i.e. Major Depressive Disorder or Dysthymia) and anxiety (i.e. generalized anxiety disorder) or individuals with only a depressive disorder.

Reliability

Internal consistency reliability and item analysis relate to how well the items are correlated with each other and are measuring the same underlying trait. Cronbach’s alpha coefficient (α) is the most commonly used measure of internal consistency reliability and is based on the pair-wise correlations between items (Cronbach, 1951). Interpretation of α is as follows: $\alpha \geq 0.90$ is considered excellent, $0.9 > \alpha \geq 0.80$ is considered good, $0.8 > \alpha \geq 0.70$ is considered acceptable, $0.70 > \alpha \geq 0.60$ is considered questionable, $0.60 > \alpha \geq 0.50$ is considered poor, and $\alpha < 0.50$ is considered unacceptable. Item analysis refers to examining each item’s correlation to

the rest of the items, correlation to the test as a whole, and calculation of what Cronbach's alpha for the scale would be if the item was not included.

The PCA and EFA were used to determine the underlying dimensionality of the data. The aim of these types of analyses are to explain the covariance among a set of variables. A PCA with parallel analysis, is based on the correlation matrix of the underlying data and graphs the eigenvalues in a screeplot. A parallel analysis simulates 100 datasets with the same sample size, number of variables, same means and variances as the true variables, but any correlation among variables is due to chance alone. This method thereby allows you to look at eigenvalues greater than what you would get by chance alone and helps in determine how many potential factors are underlying the data. The number of eigenvalues above 1, as specified in the PCA with parallel analysis, was then used to guide the EFA.

Exploratory factor analysis is a method used to specify the underlying relationship between measured variables. It is a type of internal consistency reliability, in that it examines how well the items on a scale are related to each other. The EFA was done using a mean and variance-adjusted weighted least squares estimator (WLSMV), to account for the ordered categorical response categories of the IDSS-G. EFA models were evaluated based on what made theoretical sense as well as absolute fit of the models using global fit indices. Fit indices include the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). RMSEA values lower than 0.05 and TLI/CFI values above 0.90 are indicative of good model fit (Hu & Bentler, 1998)

Test-retest reliability is concerned with the consistency of the measure across time -- would get the same score on the instrument as you did today, if the test was instead administered tomorrow? Test-retest reliability was evaluated through calculation of the Pearson correlation coefficient (*r;* Pearson, 1895) between scores on the IDSS-G at the first administration and scores on the IDSS-G from the re-interview (done by the same interviewer). Correlations of |0.7| or above are considered very strong, correlations of |0.4| to |0.69| are considered strong, |0.3| to

|0.39| are moderate, |0.2| to |0.29| are weak, and anything less than |0.2| are considered negligible (Cohen, 1988).

Inter-rater reliability relates to whether you would get the same score on the instrument if it was administered by one person compared to another person. When scores are continuous it can be measured by the intra-class correlation (ICC). The ICC has at least six different forms depending on the design of the study and the researcher's assumptions about the participants and the raters. The six main forms of the ICCs are (1,1) (2,1) (3,1), (1,k), (2,k), and (3,k) and calculated from results obtained from analysis of variance (ANOVA) of repeated measures. The first number in the ICC forms refers to rater circumstances: a 1 means that each subject is rated by a different set of *k* raters, who are randomly selected from a larger population of raters; a 2 means that a random sample of *k* raters is selected from a larger population of raters and each rater tests each subject; and a 3 means that each subject is rated by the same *k* raters who are the only raters of interest and results cannot be generalized to raters outside of the given study. The second number in the ICC forms relates whether the unit of analysis is 1 rating or relates to the mean of all measurements (*k*) (Rankin & Stokes, 1998; Shrout & Fleiss, 1979). For the IDSS-G, after the initial administration, $n = 30$ were re-interviewed by a different interviewer and ICC (1, *k*) was calculated as the measure of inter-rater reliability. Intra-class correlations greater than 0.75 are considered excellent; 0.40-0.75 are considered fair to good; and less than 0.40 considered poor (Fleiss, 1986).

In addition, because part of the instrument testing involved ratings by multiple psychiatrists, inter-rater reliability was also calculated for psychiatrist ratings using the Kappa statistic (Cohen, 1960). The Kappa statistic is the appropriate for inter-rater reliability when the data is categorical. As psychiatrists initially performed evaluations in pairs, Kappas were generated for each pair of psychiatrists separately. A Kappa of less than 0 indicates less than chance agreement; Kappa of 0.01-0.20 is indicative of slight agreement; 0.21-0.40 indicates fair

agreement; 0.41-0.60 indicates moderate agreement; 0.61-0.80 indicates substantial agreement and 0.81-0.99 indicates almost perfect agreement (Viera & Garrett, 2005).

Validity

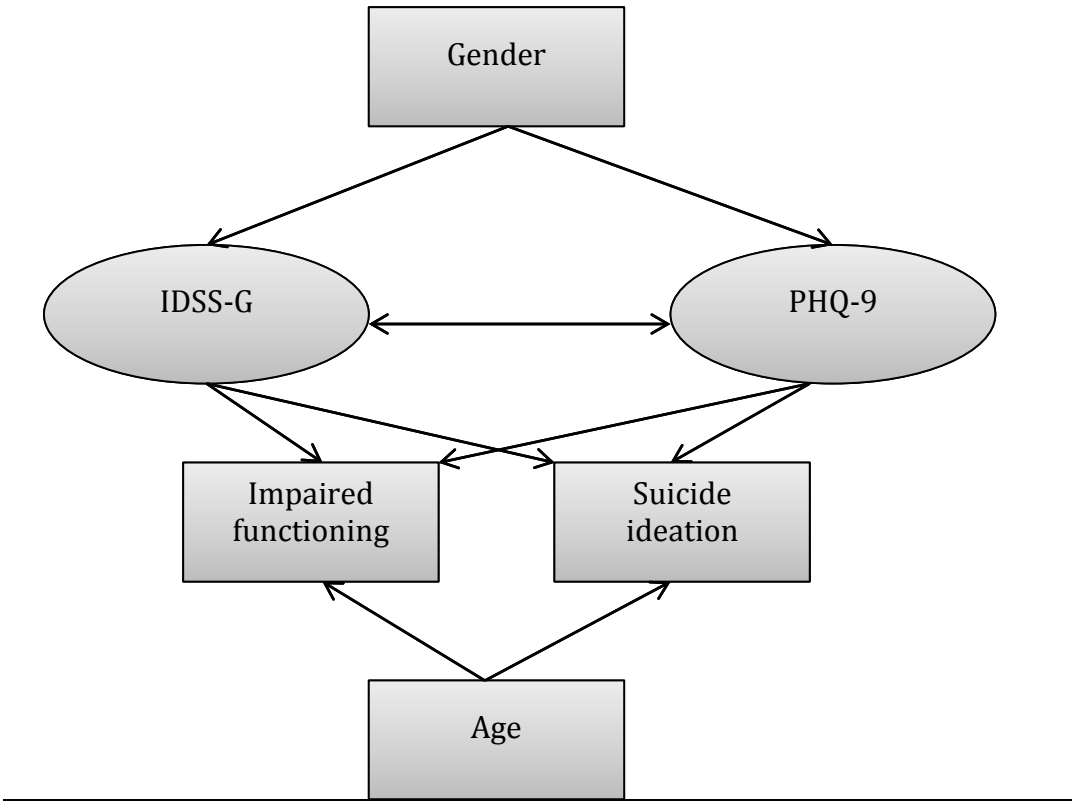
A measure is considered to have face validity when it appears to be measuring what it purports to measure (Allen & Yen, 2002). Construct validity is defined as the degree to which a scale measures the theoretical construct that it was designed to measure and is correlated to other related constructs (Allen & Yen, 2002). Construct validity can also be thought of as the extent to which an instrument fits into its nomological network. Criterion validity refers to the association of a score on a scale to a criterion variable (in this case psychiatric diagnosis) (Allen & Yen, 2002). Finally incremental validity, or the ability of the IDSS-G to increase predicative ability beyond existing measures of depression, was also evaluated (Sackett & Lievens, 2008).

Face validity was examined as part the training of the psychiatrists and during cognitive interviews. During training the psychiatrists commented on the meaning of each item and its relevance to depression. During cognitive interviews, participants described the meaning and their thought process when answering the items. For construct validity, a nomological network was developed involving the IDSS-G, the PHQ-9 (a measure of Western-defined depression), age, gender, impaired functioning, and the presence/absence of suicide ideation (Figure 6). Based on evidence in the literature, it was hypothesized that higher scores on the IDSS-G would be associated with increasing age (Bromet et al., 2011; Jorm, 2000; Kessler et al., 2003); female gender (Bromet et al., 2011; Nolen-Hoeksema, Larson, & Grayson, 1999); worse impaired functioning (Kessler & Bromet, 2013; Ormel et al., 2008), and suicidal ideation (Nock et al., 2008). In addition, high scores on the IDSS-G were hypothesized to be strongly associated with high scores on the PHQ-9. If a strong association between these two measures were present this would be considered as evidence supporting construct validity instead of evidence for criterion

validity given that the PHQ-9 had never been used in the study population and not be considered a 'gold standard.' The PHQ-9 was also thought to be measuring Western presentations of depression, which were thought to be highly associated with the IDSS-G, but not necessarily the same exact construct.

Figure 3.6

Nomological network for construct validity of the IDSS-G



Criterion validity was assessed by comparing scores on the IDSS-G to psychiatrist ratings of each participant. Each participant was evaluated by a local psychiatrist using an electronic version of the Structured Diagnostic Interview for DSM-IV (SCID; (First, Spitzer, Gibbon, & Williams, 2012). The psychiatrist classified the individual as having depression, dysthymia,

generalized anxiety disorder (GAD), or none of these disorders. Comorbidity was allowed, meaning that a person could be classified as having more than one of the disorders mentioned above.

Given that there is no dependable biomarker indicative of depression, psychiatrist ratings using the SCID, served as the "gold standard" criterion variable. A gold standard indicates "a relatively irrefutable standard that constitutes recognized and accepted evidence that a certain disease exists" (Kassirer, Kopelman, & Wong, 1991). When a new scale is developed, it is important to test this instrument against a gold standard in order to determine how well the instrument identifies cases, as well as determine the sensitivity and specificity of the instrument. Sensitivity is defined as the proportion of individuals correctly classified as having an attribute (or disorder) by the instrument compared to the gold standard. Specificity refers to the proportion of individuals correctly identified as not having the attribute (or disorder) by the instrument compared to the gold standard. This testing against a gold standard relates to the validity of the instrument and allows a determination about the extent to which the instrument measures what it purports to measure.

Reliance on an evaluation and diagnosis by a psychiatrist as a gold standard can be problematic. First, comprehensive psychiatric evaluations are often lengthy and time consuming, making them unfeasible for use in large-scale studies. Second, research has shown that agreement between self-report and clinical assessment is only moderate (Anthony et al., 1985), begging the question of which one is actually representing the truth. Finally, in much of the world, psychiatrists or trained psychological professionals are simply unavailable (Geneva, 2011), making evaluation of criterion validity using psychiatrist diagnosis, challenging. In the present study, it was determined that there were enough available local personnel to be able to use psychiatrist evaluation as the criterion. Additional, efforts were made to reduce respondent and interviewer burden, by limiting the scope of evaluation to just a few select disorders (depression, dysthymia, GAD).

For the criterion validity analysis, participants were separated into various diagnostic categories based on the psychiatrists' evaluations. The categories included: 1) no disorder vs. any disorder; 2) MDD/dysthymia vs. no disorder; 3) GAD vs. no disorder; and 4) GAD vs. MDD/Dysthymia. Criterion validity was evaluated by comparing the mean scores for each diagnostic category classification. For example, the mean score on the IDSS-G of people classified as having no disorder was compared to the mean score on the IDSS-G of people classified as having any of the disorders. Student's t-tests were used to determine if there was a statistically significant difference between the means on the IDSS-G for each diagnostic category. Criterion validity would be supported if the mean on the IDSS-G is significantly higher for those classified as having one or a combination of the disorders than for those classified as having no disorder.

Incremental validity was especially important to evaluate with the IDSS-G. As discussed previously, measurement of depression in a range of cultures and contexts has either been done through use of a western-developed measure, a locally-derived measure, or an adapted western measure with the addition of local symptoms. The IDSS reflects a different approach, which was to create a measure based on empirical evidence of what symptoms of depression look like in global populations. In order to partially justify its use, the IDSS-G would need to show incremental validity when compared to a purely western measure of depression (in this case the PHQ-9). In other words, in order to be incrementally valid, the IDSS-G would need to be better than the PHQ-9 at identifying individuals with mental health problems. To do this scores on the IDSS-G and PHQ-9 were compared to see which scale was better at predicting functioning impairment. Multiple linear regression including the IDSS-G, the PHQ-9, as well as age, and suicidal ideation was used to see if either one of the scales predicted functional impairment above and beyond the contribution of the other scale.

Clinical Utility

To explore clinical utility, receiver operating curves (ROC) were used to compare the area under the curve (AUC), for the IDSS-G across diagnostic comparisons (no disorder vs. any disorder, MDD/dysthymia vs. no disorder, GAD vs. no disorder, and GAD vs. MDD/Dysthymia). ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity), for various cutoff values of a continuous measure compared to a dichotomous criterion. An AUC of 0.5 (50% sensitivity and 50% specificity) indicates that the test is of no diagnostic utility, while an AUC of 1.0 (100% sensitivity and 100% specificity) indicates the test under investigation perfectly predicts the criterion. A general guide to interpreting AUC values is that 0.50-0.70 indicates low accuracy; 0.70-0.90 indicates moderate accuracy, and above 0.90 indicates high accuracy (Fischer, Bachmann, & Jaeschke, 2003). A cut-off point was generated for each scale based on maximizing the sensitivity and specificity for each diagnostic category (Liu, 2012).

Qualitative methods used for reliability and validity

In addition to reliability and validity testing through quantitative methods, two qualitative methods were used to assist in evaluating the reliability and validity of the IDSS-G. Pile sort activities were done with $n = 50$ participants. Participants were given 29 index cards, one with each symptom from the IDSS-G. Participants were then asked to sort the cards into piles of symptoms they thought went together. After sorting all the cards, participants were asked to name each pile and provide a reason for grouping the symptoms together in the manner they did. This activity served as a qualitative approach to factor analysis, whereby information was gathered about how participants grouped the symptoms together. The results of the pile sort can be compared to the results from the EFA, and provide qualitative support for or against the factors that emerged during the factor analysis. This strategy has been used previously in instrument

creation and validation studies in LMIC (Bolton et al., 2002; Rasmussen, Katoni, Keller, & Wilkinson, 2011).

Cognitive interviews were done with n = 60 participants (n = 30 men and n = 30 women) after completion of the IDSS-G. Cognitive interviews involved only a select number of the symptoms from the IDSS-G (13 symptoms) as some of the symptoms on the IDSS had been previously tested in a similar population (see (Haroz et al., 2014). For each symptom, participants were asked: 1) *Please describe the meaning of this question in your own words. Please use examples to help describe the meaning*; 2) *Is there any part of this question you don't understand or that does not make sense?*; 3) *Can you tell me what thought you had when deciding your answer choice? I'd like to know anything you thought of between when I asked you the question and when you gave me your answer*; and 4) *Was this question easy or difficult to answer? Could you tell me why it was difficult*?

Cognitive interviewing is a qualitative research method often used to improve questionnaires and surveys. The general theory behind this approach is that for many survey-questions participants use a number of cognitive steps, some explicit and some implicit, when deciding how to respond. The goal of the interview is to prompt the participant to reveal the thoughts they had as they answered the question. This method can be used to determine whether people understand the question, including specific words and phrases in the question; whether the question is interpreted the way it was intended to be interpreted; what types of information the participant uses to answer the question; and whether the response categories match the answer in the participant's mind (Willis, 2004).

There are two principal types of cognitive interviewing: the "think-aloud" interview and verbal probing techniques. In the "think-aloud" approach, the participant is instructed to "think aloud" as they answer each of the questions. The whole thought process is recorded and the interviewer rarely interrupts except to clarify. The verbal probing technique involves the interviewer using "probes" after the participant answers the question to further investigate their

response (Willis, 2004). For the Aim 3 study, due to the length of the IDSS-G, verbal probing was selected as the best way to do cognitive interviewing.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders fifth edition* (5th ed.*)*. Washington, DC: Author.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*(1), 37-48.

Anthony, J. C., Folstein, M., Romanoski, A. J., Von Korff, M. R., Nestadt, G. R., Chahal, R., . . . Kramer, M. (1985). Comparison of the lay diagnostic interview schedule and a standardized psychiatric diagnosis: Experience in eastern Baltimore. *Archives of General Psychiatry, 42*(7), 667-675.

Applied Mental Health Research Group (AMHR). (2013). *Design, implementation, monitoring and evaluation of cross-cultural trauma related mental health and psychosocial assistance programs: A user's manual for researchers and program implementers*. Retrieved from: http://www.jhsph.edu/research/centers-and-institutes/center-for-refugee-and-disaster-response/response_service/AMHR/dime

Bass, J., Poudyal, B., Tol, W., Murray, L., Nadison, M., & Bolton, P. (2012). A controlled trial of problem-solving counseling for war-affected adults in Aceh, Indonesia. *Social Psychiatry and Psychiatric Epidemiology, 47*(2), 279-291.

Bolton, P., Lee, C., Haroz, E. E., Murray, L., Dorsey, S., Robinson, C., . . . Bass, J. (2014a). A transdiagnostic community-based mental health treatment for comorbid disorders: Development and outcomes of a randomized controlled trial among Burmese refugees in Thailand. *PLoS Medicine, 11*(11), e1001757.

Bolton, P., Bass, J. K., Zangana, G., Kamal, T., Murray, S., Kaysen, D., . . . Rosenblum, M. (2014b). A randomized controlled trial of mental health interventions for survivors of systematic violence in Kurdistan, Northern Iraq. *BMC Psychiatry, 14*(1), 360-375.

Bolton, P., Neugebauer, R., & Ndogoni, L. (2002). Prevalence of depression in rural Rwanda based on symptom and functional criteria. *The Journal of Nervous and Mental Disease, 190*(9), 631-637.

Bolton, P., Wilk, C. M., & Ndogoni, L. (2004). Assessment of depression prevalence in rural Uganda using symptom and function criteria. *Social Psychiatry and Psychiatric Epidemiology, 39*(6), 442-447.

Cai, L., Du Toit, S., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [computer software]. *Chicago, IL: Scientific Software International*.

Cobb, A. K., & Hagemaster, J. (1987). Ten criteria for evaluating qualitative research proposals. *The Journal of Nursing Education, 26*(4), 138-143.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Crane, P. K., Belle, G. v., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241-256.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Dixon-Woods, M., Fitzpatrick, R., & Roberts, K. (2008). Including qualitative research in systematic reviews: Opportunities and problems. *Journal of Evaluation in Clinical Practice, 7*(2), 125-133.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189-199.

First, M.B., Spitzer, R.L., Gibbon, M. & Williams, J.B.W. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP).* New York: Biometrics Research, New York State Psychiatric Institute, November 2002.

Gallo, J. J., Anthony, J. C., & Muthen, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology, 49*(6), 251-264.

Gibbons, R. D., Immekus, J. C., Bock, R. D., & Gibbons, R. D. (2007). The added value of multidimensional IRT models. *Chicago: Center for Health Statistics*. Chicago Illinois: University of Illinois.

Hambleton, R. K., Waminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (1st ed.). California: Sage Publications Inc.

Harden, A., Garcia, J., Oliver, S., Rees, R., Shepherd, J., Brunton, G., & Oakley, A. (2004). Applying systematic review methods to studies of people's views: An example from public health research. *Journal of Epidemiology and Community Health, 58*(9), 794-800.

Haroz, E. E., Bass, J. K., Lee, C., Murray, L. K., Robinson, C., & Bolton, P. (2014). Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the Thailand Burma border. *BMC Psychology, 2*(1), 31-40.

Hesbacher, P. T. (1980). Psychiatric illness in family practice. *Journal of Clinical Psychiatry; Journal of Clinical Psychiatry, 41,* 6-10.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Kassirer, J. P., Kopelman, R. I., & Wong, J. B. (1991). *Learning clinical reasoning*. Baltimore, MD: Williams & Wilkins.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*(9), 1-7.

Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in Medicine, 31*(23), 2676-2686.

Mays, N., & Pope, C. (1995). Rigour and qualitative research. *BMJ: British Medical Journal, 311*(6997), 109-112.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269.

Muthén, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational and Behavioral Statistics, 10*(2), 121-132.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (Seventh Edition ed.). Los Angeles, CA: Muthén & Muthén.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58*(347-352), 240-242.

Rankin, G., & Stokes, M. (1998). Reliability of assessment tools in rehabilitation: An illustration of appropriate statistical analyses. *Clinical Rehabilitation, 12*(3), 187-199.

Rasmussen, A., Katoni, B., Keller, A. S., & Wilkinson, J. (2011). Posttraumatic idioms of distress among darfur refugees: Hozun and majnun. *Transcultural Psychiatry, 48*(4), 392-415.

Robert F. DeVellis. (2012). *Scale development: Theory and applications* (3rd ed.). California: SAGE publications, Inc.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annu.Rev.Psychol., 59*, 419-450.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R. K. Hambleton. *Handbook of Modern Item Response* Theory (85-100). New York: Springer.

Shaw, R. L., Booth, A., Sutton, A. J., Miller, T., Smith, J. A., Young, B., . . . Dixon-Woods, M. (2004). Finding qualitative research: An evaluation of search strategies. *BMC Medical Research Methodology, 4*(1), 5-10.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

StataCorp. (2013). *Stata statistical software* (Release 13 ed.). College Station, TX: StataCorp LP.

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*(11), S152-S170.

Thomas, E., & Magilvy, J. K. (2011). Qualitative rigor or research validity in qualitative research. *Journal for Specialists in Pediatric Nursing, 16*(2), 151-155.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*(1), 45-55.

Tobin, G. A., & Begley, C. M. (2004). Methodological rigour within a qualitative framework. *Journal of Advanced Nursing, 48*(4), 388-396.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care, 19*(6), 349-357.

Weiss, W., & Bolton, P. (2010). *Assessment of torture survivors in southern Iraq: Development and testing of a locally-adapted assessment instrument.* United States Agency for International Development.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, California: Sage Publications.

Winokur, A., Winokur, D. F., Rickels, K., & Cox, D. S. (1984). Symptoms of emotional distress in a family planning service: Stability over a four-week period. *The British Journal of Psychiatry, 144*(4), 395-399.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339-361.

World Health Organization. (2011). *Mental health atlas, 2011.* Geneva: World Health Organization.

Global signs and symptoms of depression: A systematic review of qualitative literature

Abstract

Background: Existing self-report scales aimed at measuring depression have mostly been developed based on western clinical populations. However, there may be signs and symptoms of depression that do not appear in western clinical populations, but are common ways of expressing depression in much of the world. **Objectives:** This review aimed to identify those signs and symptoms through a systematic review of qualitative literature related to depression. **Methods:** Nine online databases were searched for articles that related to depression and used qualitative methods. A total of 7972 articles were identified and 106 met full inclusion criteria. **Results:** These 106 studies represented data on 138 different study populations and 69 different countries/ethnicities. Depressed mood (n = 94) was the most frequently mentioned symptom across study populations. Nearly half of the top 15 most frequently mentioned symptoms (n = 7; 47%) are also symptoms that are part of the DSM-5 diagnosis of Major Depressive Disorder (MDD). Aside from DSM-5 symptoms, social isolation/loneliness (n = 70), crying a lot (n = 64) and general aches and pains (n = 54) were also commonly mentioned across all study populations. **Implications:** Findings from this review should be used to inform improvement in measurement scales and provide evidence to support the notion that depression is a universal human phenomenon, which presents relatively similarly across global populations.

Introduction

Depression is a major public health problem with an estimated point prevalence rate of 4.7% (Ferrari et al., 2012). It is a leading cause of disability, particularly among young people aged 14-44 (Murray et al. 2013). However large variations in epidemiological estimates of depression across populations have been found. For example, in two major cross-national studies of depression, Weissman et al. (1996) found lifetime prevalence ranging from 1.5% in Taiwan to 19.0% in Lebanon and Andrade et al. (2000) found lifetime prevalence rates from 1.0% in Czech Republic to 16.9% in the United States. This heterogeneity is even more pronounced in the context of war or displacement, prevalence rates have been even more varied, with point prevalence rates ranging from 3% to 85.5% (Steel et al., 2009).

Variability in these prevalence estimates may be due to substantive factors such as differences in genetic vulnerabilities or environmental risk factors. Or this heterogeneity may be due to measurement factors such as psychometric properties of the instruments and cultural differences in the meaning of the items used to measure depression (Rodin & van Ommeren, 2009). Most cross-national studies assessing the epidemiology of depression have utilized measurement instruments developed in the United States or European contexts that reflect the Western psychiatric nosology of depression reflected in the Diagnostic and Statistical Manual (American Psychiatric Association, 2013) or the International Classification of Disorders (World Health Organization, 1992). However, as the majority of the world's population lives outside of North America and Europe, the application of these instruments to non-Western populations may lack validity and reliability.

Measurement instruments of depression that are commonly used in non-Western contexts include both schedules and scales. Schedules are aimed at classifying people into diagnostic categories. Scales, by contrast, treat psychopathology as dimensional. The most commonly used schedule globally is the Composite International Diagnostic Interview (CIDI; Robins et al., 1988), which was originally developed with Western clinical populations but has been applied in

settings as varied as Ethiopia (Gelaye et al., 2013), Nepal (Ghimire, Chardoul, Kessler, Axinn, & Adhikari, 2013), Brazil (Quintana, Gastal, Jorge, Miranda, & Andreoli, 2007) and many North American and European countries (Wittchen et al., 1991). Scales, in particular self-report scales, are often more widely used, especially in low-resource settings (LRS). Many of these scales are freely available, short enough to lend themselves to easy translation and adaptation, and do not require highly trained personnel to administer. Common self-report scales for depression include the Hopkins Symptom Checklist (HSCL; Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974), the Center for Epidemiological Studies Depression (CES-D; Radloff, 1977), and the Patient Health Questionnaire (PHQ; Spitzer, Kroenke, & Williams, 1999). All of these scales were originally developed in the United States in either clinical or general populations but have been used in a wide variety of populations outside the United States (Adewuya, Ola, & Afolabi, 2006; Ertl et al., 2010; Haroz et al., 2014; Lotrakul, Sumrithe, & Saipanish, 2008; Marc et al., 2014; Silove et al., 2007).

The potential for divergence between Western developed self-report scales and signs and symptoms of depression in other contexts is acknowledged in the global mental health field (Bolton, 2001; Flaherty et al., 1988; Folmar & Palmes, 2009; Kohrt et al., 2011). However, the extent to which these discrepancies may contribute to measurement error has yet to be fully explored. Despite calls to adapt assessment instruments to local cultural contexts (Bass, Bolton, & Murray, 2007; Miller et al., 2010), it is often not done. Moreover, there is currently no single agreed upon method for comprehensive adaptation and validation of instruments (Kohrt et al., 2011). However, even when scales are adapted comprehensively, the use of Western scales as the starting point for adaptation may be problematic.

The process of applying existing measurement tools reflects a universalist epistemology. This process starts with scales that reflect Western signs and symptoms of depression and then evaluates whether these same signs and symptoms are reliable indicators of depression for people in other contexts. Universalist epistemology focuses on the universality of constructs and is

contrasted with a particularlist epistemology, in which the culturally specific aspects of constructs are emphasized. A particularlist approach would aim to capture cultural influences on depression through context specific research, but in doing so, usually lacks generalizability. Some researchers have been able to combine both universalist and particularlist approaches to inform scale adaptation (see examples: AMHR, 2013; Rasmussen et al., 2014). These methods usually start with a western-development measure (such as the HSCL) and augment the scale with context specific symptoms, to improve the scales local reliability and validity.

The problem with relying on scales originally developed in Western populations, even if adapted and tested, is that the scales are dependent on Western psychiatric nosology and the research is potentially vulnerable to *category fallacy* (Kleinman, 1977). *Category fallacy* relates to the assumption that because symptoms can be identified in different cultural contexts, they have the same meaning or constitute the same syndrome across these settings. As Kleinman states "applying such categories in non-western cultures…by definition it will find what is universal and it will systematically miss what [is not], but what is missed is more interesting…because missed symptoms will be the most striking examples of the influence of culture on depression," (Kleinman, 1977). Kleinman is correct that many of these "missed symptoms" – those symptoms of depression that are not included in Western psychiatric nosology – will be those symptoms that are influenced by non-Western culture. However, it is also plausible that important more-or-less "universal" symptoms of depression exist that for historical and/or other contingent reasons are not included in canonical Western definitions of depressions.

Taking a broader approach and examining the signs and symptoms of depression that have emerged through qualitative research all over the world not only allows for identifying the culturally specific symptoms, but has the potential to identify those near universal symptoms of depression that have not been included in western diagnostic categories. To date there has been no comprehensive systematic examination of signs and symptoms of depression that are reported

in qualitative literature (and/or from a particularlist epistemology) from a variety of cultural contexts.

The current study seeks to identify a set of signs and symptoms that have been described in qualitative research on depressive-like syndromes in a range of contexts. By reviewing this literature we hope to identify common signs and symptoms of depression that are not currently captured in Western psychiatric nosology. Reviewing qualitative literature allows a better understanding of symptoms that emerge through subjective open ended inquiry, rather than confirmed through use of established measurement instruments. This knowledge has the potential to inform and improve cross-cultural measurement of depression, by identifying common expressions of depression that could be incorporated into measurement instruments for both research and clinical use. Moreover, results from this review will contribute to a better understanding of whether depression is something we share as humans, and if so, what it looks like.

Methods

Literature Search

Qualitative literature related to depression was examined through a search of peer-reviewed academic journals and solicitation of non-peer-reviewed programmatic reports related to mental health programs in low resource settings (LRS). The literature review followed PRISMA guidelines (Moher, Liberati, Tetzlaff, & Altman, 2009) (See Appendix A for PRISMA checklist). The search was done between August 2012 and February 2013. The first step involved using a multi-step search of 8 different databases including Pubmed, Web of Science, PsycInfo, Scopus, Embase, Anthrosource, Anthropology Plus, Global Health, and Sociological Abstracts. The first step in the search involved the use of following terms: “depression,” “depressive disorder,” “melancholia,” and “depressive disorder, major.” Once those results were returned, the second step involved reviewing this subset of articles for the following terms “anthropology,”

"qualitative," "ethnography," "cross-cultural comparison," "ethnopsychology," cultural characteristics," "cross cultur*," "phenomenology," and "idioms of distress." Study titles, abstracts, and subjects were search and MeSH terms were used when possible. After this initial search, Google Scholar (up to the first 10,000 hits) was used to find any other possible references that were not found during the initial search. Finally, if a review article was identified during the search, the bibliography of the review article was also searched for possible relevant citations, and if any articles not already identified were listed in the bibliography, these were included for full-text review.

The titles and abstracts of all articles that arose during the search process were reviewed to determine if they met inclusion criteria. Inclusion criteria was as follows: 1) article must represent qualitative research; 2) article must have depression as the main focus; 3) and included information on symptoms of depression; 4) article must be written in English; and 5) and article must report on a study population of adults between the ages of 18-65. Any articles that reported data on only one individual or on a small series of case-studies was excluded due to lack of generalizability. As there are no single guidelines for the appraisal of the rigor of qualitative research (Thomas & Magilvy, 2011; Tobin & Begley, 2004), articles were not appraised on the quality of the original research.

Review and data extraction

After title and abstract review, each article that met inclusion criteria (as mentioned above) was reviewed in full and the following data was extracted by the author from all eligible articles: a) sex of the study population; b) region of the world; c) nationality and/or ethnicity; d) religious distinction if available; e) class distinction if available; f) whether the study took place in the perinatal context, or the context of war, trauma or displacement; g) whether the study took place in an urban or rural location if available; h) which qualitative research methods were used; and i) the symptoms of depression that were mentioned in the text.

Coding

Once data was extracted from all articles, the symptoms of depression were coded using a priori and emergent coding. A priori codes were based on the symptoms of Major Depressive Disorder included in the DSM-5 (American Psychiatric Association, 2013) (Table 1). Emergent coding took place as the symptoms from each article were reviewed and symptoms representing the same idea were grouped together. Multiple rounds of coding were done in order to group together all symptoms mentioned more than once. After all articles had been reviewed and respective symptoms coded, a quantitative dataset was compiled. The dataset included rows for each study population represented in the articles and columns with the name of each coded symptom. As some articles reported on multiple study populations, the number of rows in the dataset was greater than the number of articles included in the study. For each study population (row), whether or not the symptom was reported was marked as present or not present (dichotomous).

Table 4.1

A-prior codes based on DSM-5 diagnostic criteria for Major Depressive Disorder

1. Depressed mood
2. Diminished interest or pleasure
3. Significant weight loss or weight gain
4. Insomnia or hypersomnia
5. Psychomotor agitation or slowing
6. Fatigue or loss of energy
7. Worthlessness or inappropriate guilt
8. Diminished ability to think or concentrate, or indecisiveness
9. Recurrent thoughts of death/suicidal ideation
10. Functional impairment
11. Irritability

Analysis

This final dataset representing all study populations included in the review and the presence or absence of symptoms was then analyzed to examine various patterns in the data. Basic exploratory and descriptive analyses were performed to examine the most frequently mentioned symptoms and symptom variation by gender, study context, and region in the world.

Results

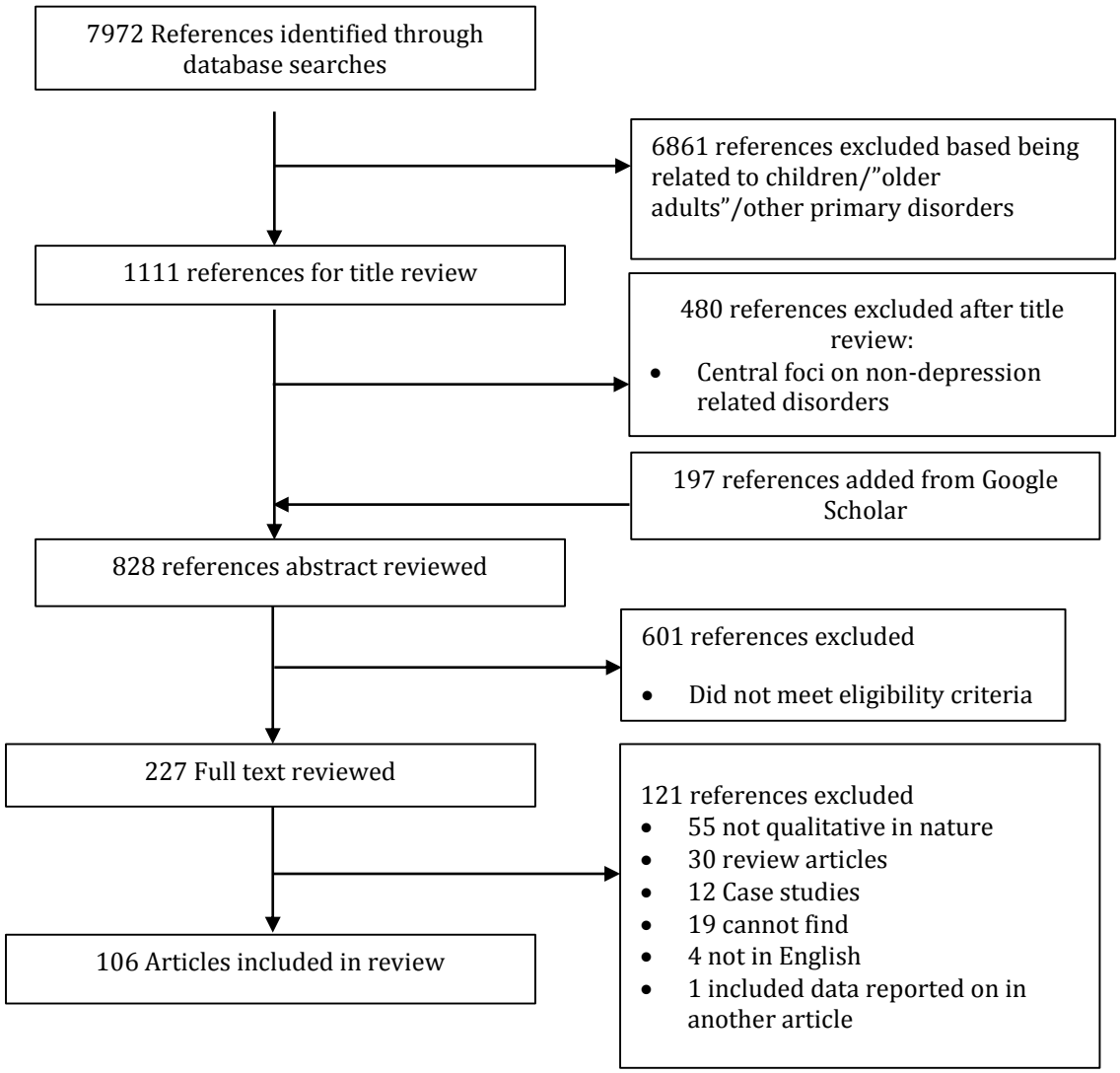
A total of 7972 references were identified through databases searches. After initial screening, 6861 references were excluded as study populations and article focus did not meet the inclusion criteria (i.e. were focused on children, older adults, or other disorders), leaving 1111 references for title review. Four hundred eighty were excluded based on title review, while 197 articles were added from the Google Scholar search, leaving 828 references for abstract review.

After abstract review, 601 articles were determined to be ineligible because of the study population did not include adults or depression was not the main focus of the article, resulting in a total of 227 references identified for full text review. One hundred and twenty-one of these references were subsequently excluded because the research did not represent qualitative research (n = 55), the article was a review (n = 30), the research represented information on a single case or multiple case studies (n = 12), the articles were not published in English (n = 4), the articles included data reported on in an earlier article (n = 1), and n = 19 could not be found through library searches. In addition, two programmatic reports that had not been published in peer-reviewed literature, were included in the review. This resulted in a total of n = 106 references that met all inclusion criteria and were included in the full review (Figure 1) (Appendix B).

Included articles described a number of methods used to elucidate signs and symptoms of depression. The most common methods were described as in-depth and/or semi-structured interviews (reported in 49 studies). Many studies used multiple methods. Focus groups were the specified methodology in n = 26 studies, and n = 16 of these, used additional qualtiative methods as well. The most common methods that were used together were interviews with key informants (i.e. local medical professionals or other locally knowledgeable individuals) and free listing exercises (n = 14). Other methods mentioned included, ethnography (n = 5), the use of the Explanatory Model Interview Catalogue (EMIC; Weiss, 1997) (n = 4), use of case vignettes (n = 3), interviews with psychiatrists (n = 2), analysis of case histories (n = 2), pile sort activities (n = 2). Four studies did not clearly specify which methods were used.

Figure 4.1

Literature review flow chart



The 106 references represent 138 different study populations and data from 69 different countries/nationalities (Figure 2). This includes data from Northern America/Europe/Australia non-native populations (35 study populations) Sub-Saharan Africa (34 study populations), South Asia (24 study populations), Latin America (21 study population), East Asia (11 study populations), Southeast Asia (9 study populations), the Middle East/North Africa (8 study

populations), and North America/Europe/Australia native populations (2 study populations). No studies were identified from Russia or Central Asia. Forty-nine study populations included data on female only samples (35.5%), 6 included data on all male samples (4.3%), and 83 included data on study samples with both males and females. Fifteen study populations focused specifically on the perinatal context (10.9%), while 21 took place in the context of war, trauma or displacement (15.0%).

Figure 4.2

Regional variation of study populations

N = 138 study populations representing 69 different nationalities and ethnicities

For coding, a priori codes included the 9 symptoms of major depression described in the DSM-V, 1 code representing the symptom of irritability and 1 code related to problems with daily functioning. A total of 80 additional emergent codes were identified during coding. For a full list of all the codes and the frequency of each code see Appendix C.

The most frequently mentioned symptoms of depression from DSM-V (i.e. from apriori codes) were *depressed mood* (*n* = 93; 67.4%), *fatigue/loss of energy* (*n* = 91; 65.9%), and *problems with sleep* (*n* = 86; 62.3%). The most frequently mentioned symptoms of depression not currently part of DSM-V diagnostic criteria (i.e. from the emergent coding) included *social isolation/loneliness* (*n* = 69; 50.0%), *crying a lot* (*n* = 64; 46.4%), *general pain* (*n* = 53; 38.4%), *anger* (*n = 50*; 36.2%) and *headaches* (*n = 49*; 36.2%) (Table 2).

Table 4.2

Top 15 most frequently mentioned symptoms across all populations combined (N = 138)

Symptom	Frequency (%)
Depressed Mood	93 (67.4)
Fatigue/loss of energy	91 (65.9)
Problems with sleep	86 (62.3)
Social isolation/loneliness	69 (50.0)
Appetite/weight	66 (47.8)
Crying a lot	64 (46.4)
Suicidal thoughts	60 (42.8)
Loss of interest	59 (43.5)
General aches and pains	53 (38.4)
Anger	50 (36.2)
Headaches	49 (36.2)
Issues with the heart	45 (33.3)
Thinking too much	44 (31.9)
Worthlessness/guilt	43 (31.9)
Hopelessness	41 (29.7)

*Bold indicates symptom included in DSM-V diagnostic criteria for Major Depression

Results by region

Depressed mood was the most frequently mentioned symptom from the DSM-V in Western non-indigenous, Middle Eastern/North African, Southeast Asian, and Sub-Saharan African populations (n = 20 [69.0%]; n = 7 [87.5%]; n = 7 [75.0%]; n = 24 [70.6%] respectively). In Latin America and East Asia the most frequently mentioned symptom was *fatigue/loss of energy* (n = 16; 76.2%; n = 10; 90.1% respectively). In South Asia the most frequently mentioned symptom was *problems with sleep* (n = 18; 75.0%). For the non-DSM

symptoms (i.e. emergent codes), the most frequently mentioned symptoms in Western non-indigenous populations were *social isolation/loneliness* (*n* = 17) and *crying* (*n* = 12); in Latin America they were *crying* (*n* = 14), *social isolation/loneliness* (*n* = 11) and *anger* (*n* = 11); in Middle East/North Africa they were *crying* (*n* = 5), *social isolation/loneliness* (*n* = 4), *general pain* (*n* = 4) and *problems with breathin*g (*n* = 4); in East Asia *general pain* (*n* = 6) and *hopelessness* (*n* = 5); in South Asia *crying* (*n* = 11), *headaches* (*n* = 11) and *issues with the heart* (*n* = 11); in Southeast Asia they were *issues with the heart* (*n* = 7), *confusion* (*n* = 5), and *disappointed* (*n* = 5); and in Sub-Saharan Africa they were *social isolation/loneliness* (*n* = 18), *headache* (*n* = 17) and *thinking too much* (*n* = 17).

In all regions, with the exception of Western Indigenous populations and Southeast Asian populations, 3 out of the top 5 most frequently mentioned symptoms are DSM-V symptoms of MDD (Table 3). Non-DSM symptoms (i.e. from the emergent coding) were only more frequently mentioned than DSM-V symptoms in two regions: Western indigenous populations and Southeast Asian populations. In studies of East Asian populations, all five top most frequently mentioned symptoms are included in DSM-5 diagnostic criteria.

Table 4.3

Top 5 most frequent symptoms by region[a]

	Western non-indigenous	Latin America	Middle East	East Asia	South Asia	Southeast Asia	Sub-Saharan Africa
1	**Depressed mood**	**Fatigue**	**Depressed mood**	**Worthlessness /guilt**	**Sleep**	Issues with heart	**Sleep**
2	Social isolation/ Loneliness	Crying	**Irritability**	**Fatigue**	**Fatigue**	Social isolation/ loneliness	**Depressed mood**
3	**Fatigue**	**Loss of interest**	Crying	**Sleep**	**Depressed mood**	**Weight/ appetite**	**Fatigue**
4	**Suicide**	**Depressed mood**	Issues with breathing	**Loss of interest**	**Weight/ Appetite**	**Depressed mood**	**Weight/ appetite**
5	Crying	Social isolation/ loneliness	**Suicidal thoughts**	**Weight/ Appetite**	Issues with heart	Confusion	Headaches

[a] Northern America/Europe/Australia Indigenous populations not included in table because of the small number of studies reported on in this region

*Bold indicates symptom included in DSM-V diagnostic criteria for Major Depression

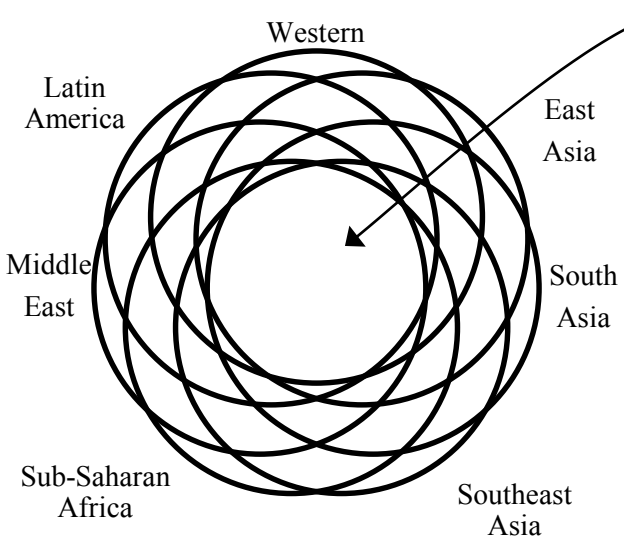
Universal signs and symptoms

Figure 3. shows all of the signs and symptoms that arose during the review and were present in all regions (with the exception of the Western indigenous populations because of the small sample size). Symptoms are ranked by their relative frequency across regions, by multiplying together their frequencies in each region. Only 20 symptoms (out of 92 symptoms)

appear in every region. Nine out of the 11 DSM-V symptoms (apriori codes) are present in every region, while the other 11 symptoms represent symptoms not currently included in DSM-5 diagnostic criteria. The DSM-V symptoms that were not present in every region were *psychomotor agitation/slowing* and *trouble concentrating*.

Figure 4.3

Universal signs and symptoms of depression



Depressed mood	Heart issues
Fatigue	**Guilt**
Sleep	**Hopeless**
Social isolation	Headaches
Weight/appetite	Thinking too much
Crying	**Irritability**
Suicidal thoughts	Anxiety
Loss of interest	Weakness
General pain	Nervous/tense
Anger	**Impaired function**

Results by gender

Among all female study populations, *fatigue/loss of energy* (n=32; 65.3%) was the most frequently mentioned symptom. Four out of the top five most frequently mentioned symptoms among women are currently part of DSM-V diagnostic criteria for MDD including *fatigue/loss of energy* (*n* = 32; 65.3%), *depressed mood* (*n* = 29; 59.2%), *problems with sleep* (*n* = 57.1%), and *weight/appetite issues* (*n* = 20; 40.8%). *Social isolation/loneliness* (*n*=20; 40.8%), *general pain* (*n* = 19; 38.8%), *headaches* (*n* = 19; 38.8%), and *crying* (n=19; 38.8%) were the most frequently mentioned symptoms not currently in the DSM-V. The number of studies involving all-male

populations was small (n = 6). The most frequently mentioned symptoms for all-male study populations included *social isolation/loneliness* (n = 5; 83.3%), *depressed mood* (n = 5; 83.3%), and *weight/appetite problems* (n = 4; 66.7%) (Table 4).

Table 4.4

Top 10 most frequently mentioned symptoms among studies of single-gender populations

Male only (n = 6)		Female only (n = 49)	
Symptom	Frequency (%)	Symptom	Frequency (%)
Social isolation/loneliness	5 (83.3)	**Fatigue**	32 (65.3)
Depressed mood	5 (83.3)	**Depressed mood**	29 (59.2)
Weight/appetite	4 (66.7)	**Sleep**	28 (57.1)
Thinking too much	3 (50.0)	**Weight/appetite**	20 (40.8)
Suicidal thoughts	3 (50.0)	Social isolation/loneliness	20 (40.8)
Loss of interest	3 (50.0)	General pain	19 (38.8)
Irritability	3 (50.0)	Headaches	19 (38.8)
Sleep	3 (50.0)	Crying	19 (38.8)
Fatigue	3 (50.0)	**Suicidal thoughts**	19 (38.8)
Crying a lot	3 (50.0)	Thinking too much	16 (32.7)

*Bold indicates symptom included in DSM-5 diagnostic criteria for Major Depression

Results by context

Results by contextual variable are presented in Table 5. In the context of trauma, most study populations reported *problems with sleep* (n = 17; 80.1%), *social isolation/loneliness* and *depressed mood* (n = 16; 76.2%), and *weight/appetite problems* (n = 15; 71.4%). In perinatal contexts, *fatigue/loss of energy* was the most common symptom (n = 11; 73.3%), followed by *social isolation/loneliness* (n = 8; 53.3%) and *crying* (n = 8; 53.3%).

Table 4.5

Top 10 most frequently mentioned symptoms in the perinatal and trauma contexts

Perinatal (n = 15)		Trauma (n = 21)	
Symptom	Frequency (%)	Symptom	Frequency (%)
Fatigue/lack of energy	11 (73.3)	**Problems with sleep**	17 (80.1)
Social isolation/loneliness	8 (53.3)	Social isolation/loneliness	16 (76.2)
Crying	8 (53.3)	**Appetite/weight**	15 (71.4)
Appetite/weight	7 (46.6)	Hopelessness	13 (61.9)
Irritability	7 (46.6)	**Fatigue/lack of energy**	13 (61.9)
Problems with sleep	7 (46.6)	Crying	12 (57.1)
Anger	7 (46.6)	**Suicidal thoughts**	11 (52.4)
Suicidal thoughts	6 (40.0)	Thinking too much	10 (47.6)
Depressed mood	6 (40.0)	**Loss of interest**	9 (42.9)
Worry & Thinking too much	5 (33.3)	Anger	9 (42.9)

*****Bold indicates symptom included in DSM-5 diagnostic criteria for Major Depression

Specific signs and symptoms related to region or context

Some symptoms that arose in the review, were mostly mentioned in some specific regions or contexts, and may reflect context specific descriptions of depression. *Substance use/abuse* was most often mentioned in Latin American study populations (n = 8; 44.4%) and was only mentioned in 5 other study populations (n=4 in Western non-indigenous and n=1 in Western indigenous). Similarly, *problems with grooming*, was mentioned the most in Latin American study populations (n = 7; 38.9%) and was only otherwise mentioned in the Middle East (n = 1) and Western non-indigenous (n = 3) study populations. *Disappointed* was rarely mentioned overall, but arose in n = 5 (55.6%) of the Southeast Asian study populations.

Aggression, was only mentioned in $n = 9$ study populations, $n = 5$ (55.6%) of which were from Sub-Saharan Africa. *Feeling suspicious*, only appeared in $n = 13$ study populations, but $n = 8$ (61.5%) of these study populations were in the context of war or displacement (See Appendix X for symptom frequencies).

Discussion

DSM-V symptoms were frequently mentioned across all regions. It seems that globally, symptoms of *depressed mood, fatigue/loss of energy, problems with sleep, appetite/weight problems, suicidal thoughts, loss of interest,* and *worthlessness/guilt* are common. Although not as frequent, *irritability* also arose in all regions. The DSM-V symptoms of *problems with concentration* and *psychomotor agitation or slowing*, were not as universally common. The evidence from this review indicate that despite the fact that DSM-5 criteria are based on Western clinical samples, these symptoms are often mentioned across contexts.

Symptoms not currently part of DSM-V diagnostic criteria, but which are commonly reported across study populations, include: *social isolation/loneliness*, *crying, general pain, headaches, anger, issues with the heart, thinking too much* and *hopelessness*. Overall, 80 non-DSM symptoms (i.e. emergent codes) arose from the data, indicating that, while DSM-V symptoms are frequently reported, there is a large variability in the way that people report signs and symptoms of depression across settings.

Although impaired functioning is critical to all DSM-V disorders, problems with daily functioning only were mentioned explicitly in 21.7% of the study populations included in this review. Perhaps depression and symptoms of depression are not directly associated with self-report impairment of functioning. In population based research, the association between depression and impaired functioning is well established in the literature in both High-Resource Countries and LRC (Cardozo et al., 2004; Langlieb & Guico-Pabia, 2010; Schneider, Baron, Davies, Bass, & Lund, 2015) The findings from this review indicate that in most places in the world people do not always make the direct connection between mental health symptoms and the

impact these symptoms have on their daily functioning. These results suggest that impaired functioning should be measured separately by other instrument or determined through clinical judgment.

We did not find major differences between genders on symptoms associated with depression, which is consistent with some recent research examining gender differences in endorsement of symptoms (Emmert-Aronson & Brown, 2014). There were some minor differences in frequency of symptoms particularly in relationship to somatic symptoms, such as *fatigue/loss of energy*, *general pain* and *headaches*, which were found to be more frequently reported in female-only populations compared to male-only populations. These findings are consistent with previous findings in the United States that indicate women tend to report *fatigue/loss of energy* and other somatic complaints more than men (Khan, Gardner, Prescott, & Kendler, 2002; Marcus et al., 2005). However, evidence for differences should only be considered preliminary as the comparison of symptoms reported between genders was limited by the lack of research done in all male populations.

Notably, the symptom of *Anger* was frequently mentioned overall (n = 50; 36.2%), and was present in female-only populations (n = 13; 26.5%) and male-only populations (n = 2; 33.3%). There is general consensus that men's depression usually consists of symptoms of anger, impoverished social relationships, emotional numbness, impulse control difficulties, irritability, aggression, substance use, and suicide (Brownhill, Wilhelm, Barclay, & Schmied, 2005; Cochran & Rabinowitz, 2003; Martin, Neighbors, & Griffith, 2013; Oliffe & Phillips, 2008). However, results from this review indicate that globally women also report *anger* as a symptom of depression, which is consistent with some literature (Rees et al., 2013; Williamson, O'Hara, Stuart, Hart, & Watson, 2014). While there may be a masculine form of depression characterized by externalizing symptoms, these symptoms should not be ignored in women, and in fact may be common signs of depression in women that have not been fully recognized.

Somatic complaints were very common among all study populations. This contradicts a commonly held belief that non-Westerners are more likely to express their distress through somatic symptoms compared to Western populations (Kleinman, 2004). However, reviews of other research has shown that somatic complaints related to depression are ubiquitous (Draguns & Tanaka-Matsumi, 2003) and their expression often functions as a reflection of both the individual and the broader healthcare system (Kirmayer, 2001).

Some symptoms of depression have been shown to be culturally variant in previous literature. In their review of psychopathology across and within cultures, Draguns & Tanaka-Matsumi (2003) found *guilt* to be culturally variant and less common in some regions of the world. Our findings suggest that feelings of *worthlessness/guilt* were identified in just over one third of the study populations (n = 45; 31.9%), and were mentioned in every region (with the exception of Western indigenous populations). It was the most commonly mentioned symptom in East Asia (n = 10; 83.3%), but was relatively rarely mentioned in Latin American study populations (n = 3; 14.3%). These results suggest that while *worthlessness/guilt* may be rather uncommon in some populations, this symptom is present across genders and contexts.

Other symptoms seemed to be contextually specific and should be considered for inclusion in measurement tools used among these specific populations. For example, in study populations from Latin America, common signs and symptoms of depression included substance abuse and problems with grooming. The relatively high frequencies of these symptoms in Latin America could represent the importance of these symptoms in expression of depression in settings in this region. This is not to say that these two symptoms are not important in other regions, but perhaps in Latin America people are more likely to endorse these items if they are depressed and thus should be considered for inclusion in measurement instruments for this region.

There were many symptoms that overlapped with symptoms of anxiety disorders, such as worry, issues with breathing, irritability, problems with sleep, and restlessness. While we limited our search strategy to articles with a main focus of depression, we did not exclude articles that

focused on anxiety (as long as depression was one of the main foci of the article). It is also likely that some articles reported on individuals with comorbid depression and anxiety disorders, resulting in reports of symptoms related to both categories of psychopathology. Disentangling the distinction between depression and anxiety symptoms was not possible given the diversity of methods used and reported on in the articles. Moreover, there is a multitude of evidence showing that depression and anxiety are often comorbid (Kessler et al., 2008), share similar risk factors (Almeida et al., 2012), have similar neurocognitive processes involving the limbic system (Ressler & Nemeroff, 2000), and respond to similar treatments (Butler, Chapman, Forman, & Beck, 2006). Our review suggests that the signs and symptoms of depression and anxiety globally, are quite intertwined (Abas & Broadhead, 1997; Bener, Ghuloum, & Abou-Saleh, 2012; Das-Munshi et al., 2008; Kaaya et al., 2002).

Results from this review could be used as the foundation for an item bank of symptoms related to depression for the creation of self-report scales that have potential global applicability. In the United States, the National Institute of Health has created the Patient-Reported Outcomes Measurement Information System (PROMIS®) which is a computer adapted measurement instrument aimed at assessing emotional distress, pain, fatigue, sleep disturbance, physical functioning and social participation (for detailed information, see www.nihpromis.org). Items used for measuring depression were assembled by way of a comprehensive literature search to identify quantitative measures of depression. The PROMIS literature search resulted in 1204 abstracts related to depression which translated into 78 measurement scales. The items from these scales were then refined through both quantitative and qualitative methods to generate a final depression item bank of 28 items for depression (Pilkonis et al., 2011). However, the literature review for PROMIS excludes qualitative studies, which provide a valuable source of information on signs and symptoms of depression that are not captured in standard measurement tools. Perhaps, using the results from the current review in combination with the quantitative

approaches used in PROMIS®, it would be possible to create an item bank of well performing items related to depression for use in global mental health settings.

Limitations

This review has several important limitations. First, all of the articles included in this review are in English and thus the symptoms extracted have been translated into English by the authors of the studies. It is very possible that during translation, nuances of the literal expressions of the signs and symptoms, were not captured. Moreover, by limiting articles to those in English, the results may be influenced by an inherent bias in favor of DSM symptoms. English speakers seeking to better understand depression or find case examples of depressed individuals in other cultures may be primed to look specifically for people who are reporting symptoms recognized in western psychiatric nosology.

Only two articles from the non-peer reviewed grey literature were included. Thus, while we identified signs and symptoms from a comprehensive list of published literature, we may be underestimating frequency and/or missing signs and symptoms. However, we do not claim to be capturing all possible signs and symptoms related to depression. Only that if a symptom was mentioned in a study population, than it should be considered as a possible symptom of depression. If a symptom was not mentioned in one of the included study populations and did not arise in this review, it may still be very important. More qualitative research needs to be done to explore expressions of depression in other populations, particularly among males and in Central Asian study locations, which were under represented in the current study.

Another limitation is our exclusion of case studies. Some case studies can actually reflect a rich description of depression that would not have otherwise come out during broader qualitative work. However, due to the nature of this review, which was to quantify signs and symptoms of depression that have arisen in qualitative work, the rich descriptions were not the

focus. The coding of the articles was only done by one person, and thus the reliability of the coding is unknown.

We did not assess the quality of the qualitative research. While there exist some criteria for evaluating the quality of qualitative research (e.g. Lincoln & Guba, 1985) there is no consensus in the field or data supporting one approach over the other. As we wanted to err on the side of being overly inclusive, we chose to include all qualitative literature that met the inclusion criteria.

Conclusions

The aim of this review was to identify common signs and symptoms of depression that are not currently captured in Western depression nosology. By doing so, we hoped to examine both the universal and the context specific of signs and symptoms of depression in order to allow for informed improvement of measurement instruments for use in global mental health settings. Findings from this review (that similar depressive symptoms are mentioned spontaneously across diverse contexts) appear to support the claim that depression is a universal human phenomenon that presents fairly similarly across global populations.

This study is the first of its kind to systematically review signs and symptoms of depression that have emerged through qualitative research across the world. The field of global mental health relies on accurate measurement of depression for policy, research and clinical decisions. However, most measurement instruments were developed in Western populations. This review of qualitative literature was aimed at identifying common signs and symptoms of depression across genders and regions, as well as in perinatal and trauma-affected populations, and provides support for revision of measurement instruments used in global mental health settings.

References

Abas, M. A., & Broadhead, J. C. (1997). Depression and anxiety among women in an urban setting in zimbabwe. *Psychological Medicine, 27*(01), 59-71.

Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst nigerian university students. *Journal of Affective Disorders, 96*(1-2), 89-93.

Almeida, O. P., Draper, B., Pirkis, J., Snowdon, J., Lautenschlager, N. T., Byrne, G., . . . Flicker, L. (2012). Anxiety, depression, and comorbid anxiety and depression: Risk factors and outcome over two years. *International Psychogeriatrics, 24*(10), 1622-1632.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders fifth edition* (5th ed.*)*. Washington, DC: Author.

Andrade, L., Caraveo-Anduaga, J., Berglund, P., Bijl, R., Kessler, R., Demler, O., . . . Epidem, W. I. C. P. (2000). Cross-national comparisons of the prevalences and correlates of mental disorders. *Bulletin of the World Health Organization, 78*(4), 413-426.

Applied Mental Health Research Group (AMHR). (2013). *Design, implementation, monitoring and evaluation of cross-cultural trauma related mental health and psychosocial assistance programs: A user's manual for researchers and program implementers*. Retrieved from http://www.jhsph.edu/research/centers-and-institutes/center-for-refugee-and-disaster-response/response_service/AMHR/dime

Bass, J. K., Bolton, P. A., & Murray, L. K. (2007). Do not forget culture when studying mental health. *The Lancet, 370*(9591), 918-919.

Bener, A., Ghuloum, S., & Abou-Saleh, M. T. (2012). Prevalence, symptom patterns and comorbidity of anxiety and depressive disorders in primary care in qatar. *Social Psychiatry and Psychiatric Epidemiology, 47*(3), 439-446.

Bolton, P. (2001). Local perceptions of the mental health effects of the rwandan genocide. *The Journal of Nervous and Mental Disease, 189*(4), 243-248.

Brownhill, S., Wilhelm, K., Barclay, L., & Schmied, V. (2005). 'Big build': Hidden depression in men. *Australian and New Zealand Journal of Psychiatry, 39*(10), 921-931.

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26*(1), 17-31.

Cardozo, B. L., Bilukha, O. O., Crawford, C. A. G., Shaikh, I., Wolfe, M. I., Gerber, M. L., & Anderson, M. (2004). Mental health, social functioning, and disability in postwar Afghanistan. *JAMA, 292*(5), 575-584.

Cochran, S. V., & Rabinowitz, F. E. (2003). Gender-sensitive recommendations for assessment and treatment of depression in men. *Professional Psychology: Research and Practice, 34*(2), 132-140.

Das-Munshi, J., Goldberg, D., Bebbington, P. E., Bhugra, D. K., Brugha, T. S., Dewey, M. E., . . . Prince, M. (2008). Public health significance of mixed anxiety and depression: Beyond current classification. *The British Journal of Psychiatry, 192*(3), 171-177.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1974). The hopkins symptom checklist (HSCL): A self-report symptom inventory. *Behavioral Science, 19*(1), 1-15.

Draguns, J., & Tanaka-Matsumi, J. (2003). Assessment of psychopathology across and within cultures: Issues and findings. *Behaviour Research and Therapy, 41*(7), 755-776.

Emmert-Aronson, B. O., & Brown, T. A. (2014). An IRT analysis of the symptoms of major depressive disorder. *Assessment,* 1-9.

Ertl, V., Pfeiffer, A., Saile, R., Schauer, E., Elbert, T., & Neuner, F. (2010). Validation of a mental health assessment in an african conflict population. *Psychological Assessment, 22*(2), 318-324.

Ferrari, A., Somerville, A., Baxter, A., Norman, R., Patten, S., Vos, T., & Whiteford, H. (2012). Global variation in the prevalence and incidence of major depressive disorder: A systematic review of the epidemiological literature. *Psychological Medicine,* , 1-11.

Flaherty, J. A., Gaviria, F. M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J. A., & Birz, S. (1988). Developing instruments for cross-cultural psychiatric research. *The Journal of Nervous and Mental Disease, 176*(5), 260-263.

Folmar, S., & Palmes, G. K. (2009). Cross-cultural psychiatry in the field: Collaborating with anthropology. *Journal of the American Academy of Child & Adolescent Psychiatry, 48*(9), 873-876.

Gelaye, B., Williams, M. A., Lemma, S., Deyessa, N., Bahretibeb, Y., Shibre, T., . . . Vander Stoep, A. (2013). Diagnostic validity of the composite international diagnostic interview (CIDI) depression module in an east African population. *The International Journal of Psychiatry in Medicine, 46*(4), 387-405.

Ghimire, D. J., Chardoul, S., Kessler, R. C., Axinn, W. G., & Adhikari, B. P. (2013). Modifying and validating the composite international diagnostic interview (CIDI) for use in Nepal. *International Journal of Methods in Psychiatric Research, 22*(1), 71-81.

Haroz, E. E., Bass, J. K., Lee, C., Murray, L. K., Robinson, C., & Bolton, P. (2014). Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the Thailand Burma border. *BMC Psychology, 2*(1), 31-40.

Kaaya, S. F., Fawzi, M., Mbwambo, J., Lee, B., Msamanga, G. I., & Fawzi, W. (2002). Validity of the hopkins symptom Checklist- 25 amongst HIV- positive pregnant women in Tanzania. *Acta Psychiatrica Scandinavica, 106*(1), 9-19.

Kessler, R. C., Gruber, M., Hettema, J. M., Hwang, I., Sampson, N., & Yonkers, K. A. (2008). Co-morbid major depression and generalized anxiety disorders in the national comorbidity survey follow-up. *Psychological Medicine, 38*(03), 365-374.

Khan, A. A., Gardner, C. O., Prescott, C. A., & Kendler, K. S. (2002). Gender differences in the symptoms of major depression in opposite-sex dizygotic twin pairs. *American Journal of Psychiatry, 159*(8), 1427-1429.

Kirmayer, L. (2001). Cultural/variations in the clinical presentation of depression and anxiety: Implications for diagnosis and treatment. *Journal of Clinical Psychiatry, 62*(13), 22-30.

Kleinman, A. (2004). Culture and depression. *New England Journal of Medicine, 351*(10), 951-953.

Kleinman, A. M. (1977). Depression, somatization and the "new cross-cultural psychiatry". *Social Science & Medicine (1967), 11*(1), 3-9.

Kohrt, B. A., Jordans, M. J. D., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). Validation of cross-cultural child mental health and psychosocial research instruments: Adapting the depression self-rating scale and child PTSD symptom scale in nepal. *BMC Psychiatry, 11-27.*

Langlieb, A. M., & Guico-Pabia, C. J. (2010). Beyond symptomatic improvement:Assessing real-world outcomes in patients with major depressive disorder. *Primary Care Companion to the Journal of Clinical Psychiatry, 12*(2), PCC.09r00826.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry.* Newbury Park, CA: Sage Publications.

Lotrakul, M., Sumrithe, S., & Saipanish, R. (2008). Reliability and validity of the thai version of the PHQ-9. *BMC Psychiatry, 8*, 46-53.

Marc, L. G., Henderson, W. R., Desrosiers, A., Testa, M. A., Jean, S. E., & Akom, E. E. (2014). Reliability and validity of the Haitian creole PHQ-9. *Journal of General Internal Medicine, 29*(12), 1679-1686.

Marcus, S. M., Young, E. A., Kerber, K. B., Kornstein, S., Farabaugh, A. H., Mitchell, J., . . . Rush, A. J. (2005). Gender differences in depression: Findings from the STAR* D study. *Journal of Affective Disorders, 87*(2), 141-150.

Martin, L. A., Neighbors, H. W., & Griffith, D. M. (2013). The experience of symptoms of depression in men vs women: Analysis of the national comorbidity survey replication. *JAMA Psychiatry, 70*(10), 1100-1106.

Miller, K. E., Omidian, P., Quraishy, A. S., Quraishy, N., Nasiry, M. N., Nasiry, S., . . . Yaqubi, A. A. (2010). The Afghan symptom checklist: A culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry, 76*(4), 423-433.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269.

Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., . . . Abdalla, S. (2013). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010. *The Lancet, 380*(9859), 2197-2223

Oliffe, J. L., & Phillips, M. J. (2008). Men, depression and masculinities: A review and recommendations. *Journal of Men's Health, 5*(3), 194-202.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS(R)): Depression, anxiety, and anger. *Assessment, 18*(3), 263-283.

Quintana, M. I., Gastal, F. L., Jorge, M. R., Miranda, C. T., & Andreoli, S. B. (2007). Validity and limitations of the Brazilian version of the composite international diagnostic interview (CIDI 2.1). *Revista Brasileira De Psiquiatria, 29*(1), 18-22.

Radloff, L. S. (1977). The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Rasmussen, A., Eustache, E., Raviola, G., Kaiser, B., Grelotti, D. J., & Belkin, G. S. (2014). Development and validation of a haitian creole screening instrument for depression. *Transcultural Psychiatry, 52*(1), 33-57.

Rees, S., Silove, D., Verdial, T., Tam, N., Savio, E., Fonseca, Z., . . . Tay, K. (2013). Intermittent explosive disorder amongst women in conflict affected Timor-Leste: Associations with human rights trauma, ongoing violence, poverty, and injustice. *PloS One, 8*(8), e69207.

Ressler, K. J., & Nemeroff, C. B. (2000). Role of serotonergic and noradrenergic systems in the pathophysiology of depression and anxiety disorders. *Depression and Anxiety, 12*(S1), 2-19.

Robins, L. N., Wing, J., Wittchen, H. U., Helzer, J. E., Babor, T. F., Burke, J., . . . Regier, D. A. (1988). The composite international diagnostic interview: An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Archives of General Psychiatry, 45*(12), 1069-1077.

Rodin, D., & van Ommeren, M. (2009). Commentary: Explaining enormous variations in rates of disorder in trauma-focused psychiatric epidemiology after major emergencies. *International Journal of Epidemiology, 38*(4), 1045-1048.

Schneider, M., Baron, E., Davies, T., Bass, J., & Lund, C. (2015). Making assessment locally relevant: Measuring functioning for maternal depression in Khayelitsha, Cape Town. *Social Psychiatry and Psychiatric Epidemiology,* 1-10.

Silove, D., Manicavasagar, V., Mollica, R., Thai, M., Khiek, D., Lavelle, J., & Tor, S. (2007). Screening for depression and PTSD in a cambodian population unaffected by war. *The Journal of Nervous and Mental Disease, 195*(2), 152-157.

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1999). Validation and utility of a self-report version of PRIME-MD. *JAMA, 282*(18), 1737-1744.

Steel, Z., Chey, T., Silove, D., Marnane, C., Bryant, R. A. & Ommeren, M. v. (2009). Association of torture and other potentially traumatic events with mental health outcomes among populations exposed to mass conflict and displacement: A systematic review and meta-analysis. *JAMA, 302*(5), 537-549.

Thomas, E., & Magilvy, J. K. (2011). Qualitative rigor or research validity in qualitative research. *Journal for Specialists in Pediatric Nursing, 16*(2), 151-155.

Tobin, G. A., & Begley, C. M. (2004). Methodological rigour within a qualitative framework. *Journal of Advanced Nursing, 48*(4), 388-396.

Weiss, M. (1997). Explanatory Model Interview Catalogue (EMIC): framework for comparative study of illness. *Transcultural psychiatry*, *34*(2), 235-263.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H. G., . . . Lellouch, J. (1996). Cross-national epidemiology of major depression and bipolar disorder. *JAMA, 276*(4), 293-299.

Williamson, J. A., O'Hara, M. W., Stuart, S., Hart, K. J., & Watson, D. (2015). Assessment of postpartum depressive symptoms: The importance of somatic symptoms and irritability. *Assessment*, *22*(3), 309-318.

Wittchen, H. U., Robins, L. N., Cottler, L. B., Sartorius, N., Burke, J. D., & Regier, D. (1991). Cross-cultural feasibility, reliability and sources of variance of the composite international diagnostic interview (CIDI). the multicentre WHO/ADAMHA field trials. *The British Journal of Psychiatry, 159*, 645-653.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Depression symptoms across settings: An IRT analysis of the Hopkins Symptom Checklist for depression using data from eight studies

Abstract

There is considerable heterogeneity in estimates of the prevalence of depression across low and middle-income countries (LMIC). One source of this heterogeneity may be due to measurement error. An analytic method that can be used to identify sources of measurement error and adjust for potential response biases is Item Response Theory (IRT). This study involved an IRT-based analysis of data from adults in eight different settings within LMIC: Colombia, Indonesia, Kurdistan/Northern Iraq (Dohuk and Erbil/Sulaimaniya), Rwanda, Southern Iraq, Thailand (Burmese refugees), and Uganda, to understand how items on the Hopkins Symptom Checklist 15-item depression scale (HSCL) performed within and across settings. The IRT analyses provided information on discrimination parameters (*a*), location parameters (*b*), and response bias (DIF) for each item. Results showed that most items performed well across settings. The item "lost of sexual interest or pleasure" had low discrimination parameters (a = 0.74 in Rwanda to a = 1.26 in Dohuk) across settings, indicating that this item is not highly related to depression in these settings. The item related to suicidal ideation, also performed poorly. The item "blaming oneself," had low discrimination parameters in some settings (Colombia: a = 1.10; Uganda: a = 0.63) and had similar location parameters as other items, indicating possible redundancy. All items showed some degree of response bias (DIF), but bias at the item level only impacted aggregate scale-level scores in Colombia, Indonesia, Southern Iraq, and Uganda. The HSCL appears to perform well across settings, but some revisions are needed to improve its measurement properties. The results from this study will support the development of an instrument to measure depression in populations across the globe.

Introduction

Major depression has been identified as a significant contributor to the global burden of disease (Whiteford, Ferrari, Degenhardt, Feigin, & Vos, 2015). However, there is considerable heterogeneity in prevalence estimates of depression across different geographic regions. In one of the early cross national studies on Major Depressive Disorder (MDD), Weissman et al. (1996), and colleagues identified 12-month prevalence rates that ranged from 0.8% in Taiwan to 5.8% in New Zealand. Given the differing methodologies used in each of the study locations, the nature of the variation in prevalence rates from this study was particularly unclear. More recently, the World Mental Health Survey (WMHS) Consortium conducted population surveys using a standard protocol and assessment instrument. Despite standardize methods, heterogeneity in prevalence estimates persisted, with 12-month prevalence rates of MDD ranging from 2.2% in Japan to 10.4% in Brazil (Moussavi et al., 2007).

The variability in these estimates may reflect true differences in the epidemiology of depression, as explained by biological or social causes. However, it may be that measurement error is driving some of the heterogeneity. Distinguishing between true differences and artifactual differences (i.e. caused by measurement error) has significant implications for policy, program planning and evaluation, as well as clinical decision making, especially in low resource settings. At the policy level, decisions for allocation of financial resources are often influenced by epidemiologic data indicating population need. Potential over- or under-estimation of disease burden may divert limited resources from those who most need help. In terms of program planning and evaluation, the use of inappropriate measurement instruments may result in inappropriate or even harmful intervention programs (Wessells, 2009). Finally, if measurement instruments do not accurately reflect the problems people are experiencing, there is potential to miss people who may need and benefit from services or to offer services to those who would be unlikely to benefit from them. This is especially important in settings where there are shortages of

trained mental health professionals and patients rely on lay workers to assess and treat mental health problems.

One source of measurement error may be that instruments that accurately measure depression in some populations do not measure depression well in other populations. This could occur if factors other than the presence or absence of depression cause individuals in different contexts to respond differently to the items used to assess depression. For example, individuals from one cultural context may be more likely to disclose personal information than individuals from another, resulting in higher scores on a depression measure due to culture, but not reflecting actual true differences in prevalence of the underlying mental health problem. Or, it may be that there are differences in meaning of items for one group compared to another.

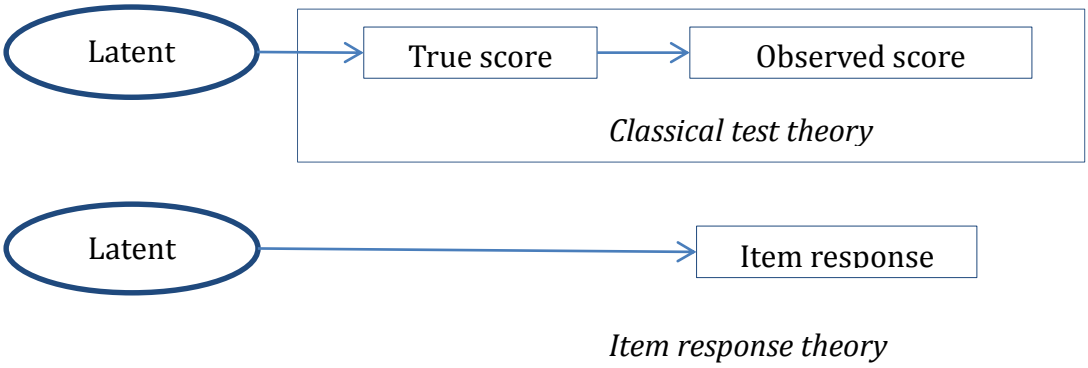
Simply taking a measurement instrument and using it across settings, without evidence to support similarity in psychometric properties in each setting, does not allow for direct comparisons of scores. Most measurement instruments that have been used in cross-cultural and global mental health research were originally developed in Western, clinical populations using Classical Test Theory (CTT). Scales based on CTT have several properties that may make use of these scales in different settings problematic. First, item and scale properties are sample dependent, meaning that scales validated in one sample (i.e. population or group of people) need to be revalidated when used in another sample. Second, to compare scores across settings, the scales have to be totally equivalent, which means all items are similarly related to the characteristic being measured. Full equivalency is hard to demonstrate. Third, CTT assumes that error is constant across all levels of severity, which may not always be the case. For example, it may be that a scale reliably measures only mild depression and is less precise in measuring more severe depression. In CTT, this scale would be assumed to reliably measure depression for all levels of severity (Embretson, 1996).

Recently, scale developers have used Item Response Theory (IRT) as a basis for scale development and refinement. IRT is a type of latent variable model, called a latent trait analysis.

Latent trait analyses are used in situations where the latent variable (i.e. unobserved variable) is thought to be continuous, but observed indicators are categorical. In IRT, the level of the latent trait (θ) directly predicts responses to items (Figure 1). In other words, IRT utilizes a conditional probability framework such that as the level of the latent trait (θ) increases the probability of endorsing the item also increases. For example, a person with severe depression would be more likely to endorse depression related items, than a person with only mild depression.

Figure 5.1

Classical Test Theory vs. Item Response Theory



For instrument development and refinement purposes, IRT has several advantages over CTT. IRT is a more complex model than CTT, but it also allows for examination of a number of individual item-level characteristics that can be used to determine the quality of existing items or the necessity of additional items. For measures of psychological latent traits, such as depression IRT models usually involve estimation of two parameters: item discrimination parameters (*a*) and item location parameters (*b*). Item discrimination parameters indicate how strongly an item is associated to the underlying latent trait (i.e. depression). Item location parameters (*b*), model the degree or severity of an underlying latent trait (i.e. depression) needed to endorse an item with a certain response (i.e. frequency of experiencing the symptom over the previous 2 weeks). In

addition, reliability (or "information" in IRT terms) can be estimated for each item independently and for the scale as a whole. Information represents the certainty to which an item or scale measures the underlying latent trait (θ) and can vary as function of the level of θ. IRT can also be used for investigating potential response bias for each item (called Differential Item Functioning or DIF). An item shows DIF if people with the same level of θ (i.e. same level of depression severity) do not have the same probability of endorsing the item (Hambleton et al., 1991).

In the context of instrument development and refinement for measuring depression across settings, examination of item parameters, item and scale information, and evaluation of DIF, allows for a better understanding of how a measure of depression performs in different populations. Parameter estimates can be used to decide which items are most relevant for a given population and which items measure less severe or more severe depression for different populations. Item information indicates the level of the latent trait where items are the most reliable for different populations. Identifying DIF may be helpful for determining whether items perform comparably across different populations. If DIF is identified, severity scores can be adjusted for these differences making direct comparisons across populations more meaningful. Examining item-level characteristics across different settings can inform instrument development and/or refinement as well as identify, possible sources of error contributing to heterogeneity in prevalence estimates.

The present study examined the performance of a measurement instrument commonly used in Low and Middle Income Countries (LMIC) to measure depression: the Hopkins Symptom Checklist 15-item depression scale (HSCL-15; Mollica, McDonald, Massagli, & Silove, 2004). The specific aims were to 1) evaluate the IRT based parameters of the HSCL-15 across and within each study population (referred to as "setting" herein); and 2) evaluate the items on the HSCL-15 for DIF across settings. The analysis draws on data from adults (over 18 years old) in eight different settings within LMIC: Colombia, Indonesia, Kurdistan/Northern Iraq (broken down into two distinct cultural settings: Dohuk and Erbil/Sulaimaniya), Rwanda, Southern Iraq,

Thailand (Burmese refugees), and Uganda. The results from this study will inform the development of a global instrument to measure signs and symptoms of depression in LMIC.

Methods

Data

The analysis was conducted using datasets from eight distinct settings, including Colombia, Indonesia, Kurdistan/Northern Iraq (Dohuk and Erbil/Sulaimaniya), Rwanda, Southern Iraq, Thailand (Burmese refugees), and Uganda. The combined data represents individual level-depression data on a total of $N = 4732$ participants (Table 1). Data from the eight datasets were combined into one dataset that included basic demographic information such as age, gender and education level, as well as the 15 items from the HSCL-15.

Table 5.1

Description of data included in IRT analysis

Study Setting	***Type of study***	***N***
Colombia*	Screening, Validity	1263
Northern Iraq/Kurdistan		
Dohuk	Clinical Monitoring	294
Erbil and Sulaimaniya	Clinical Monitoring	680
Indonesia*	Screening, Validity	588
Southern Iraq	Validity	149
Rwanda	Epidemiologic study	368
Thailand*	Screening, Validity	803
Uganda	Epidemiologic study	587
TOTAL		4732

*Observations missing all item-level data for depression were dropped from this table and all subsequent analyses (Colombia: $n = 1$; Indonesia: $n = 1$; Thailand: $n = 15$)

All data came from research conducted by the Applied Mental Health Research (AMHR) group at Johns Hopkins University and followed the same basic approach to data collection (Design, Implementation, Monitoring and Evaluation or DIME approach). Briefly, the DIME approach involves an initial qualitative study to understand relevant problems faced by a specific study population. This qualitative research goes on to inform instrument development and selection of an appropriate intervention to help address the identified mental health problems. Mental health instruments developed as part of the DIME process are tested and validated in each context and then used as screening measures for determining intervention eligibility and intervention impact (Applied Mental Health Research Group, 2013). For the current study, secondary data analysis of this existing data was approved by the JHU IRB (IRB # 4721).

All data came from trauma-affected populations. The sample from Colombia was collected as part of a validation study to test the reliability and validity of an adapted instrument and as screening for a Randomized Control Trial (RCT) of a psychotherapeutic intervention. The sample from Indonesia comes from a study validating an instrument used to assess psychological symptoms and a screening program for a randomized control trial of a psychotherapeutic intervention among torture-affected adults in Ache, Indonesia (Bass et al., 2012). The samples from Kurdistan (Dohuk, and Erbil/Sulaymaniya) are from a general clinic-based monitoring system that was established for the implementation of an RCT of psychotherapeutic interventions. The sample includes participants who were eligible for the RCT as well as all clients who were assessed at the clinic but found not to be eligible for the trial (Bolton et al., 2014b). The Rwandan data comes from a population based survey conducted of a trauma-affected adult in 5 sectors in the rural communities of Kanzenze and Butamwa in Rwanda (Bolton, Neugebauer, & Ndogoni, 2002). The Southern Iraq data are from a study that tested the reliability and validity of an instrument designed to measure psychological distress among victims of torture living in Southern Iraq (Weiss & Bolton, 2010). The data from Thailand is from completed studies aimed at testing the psychometric properties of a locally-adapted mental health assessment tool and as

screening for a RCT of a psychotherapeutic intervention among Burmese living on the Thai/Burma border who had experienced torture or trauma (Bolton et al., 2014a; Haroz et al., 2014). Finally, the data from Uganda comes from a clustered-based random survey of adults in the Rakai and Masaka districts in southwest Uganda (Bolton, Wilk, & Ndogoni, 2004).

Measurement Instrument

The HSCL-15 was administered to all participants across studies. The HSCL-15 was originally developed using a sample from an American clinical population, but is one of the most commonly used measures of depression in global mental health research (Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974; Hesbacher, 1980; Mollica et al., 2004). Respondents were asked how much each symptom (i.e. item on the scale) bothered or distressed him/her in the past couple of weeks. Response options ranged from 0 "not at all" to 3 "extremely." The timeframe varied by setting and was 2 weeks for the data from Indonesia, Kurdistan (Dohuk and Erbil/Sulaimaniya), Rwanda and Southern Iraq, 4 weeks for the data from Colombia and Thailand, and 1 week for the data from Uganda. The HSCL-15 was tested and validated in each of the eight settings (Bass et al., 2012; Bolton et al., 2004; Bolton et al., 2002; Bolton et al., 2014; Haroz et al., 2014; Weiss & Bolton, 2010).

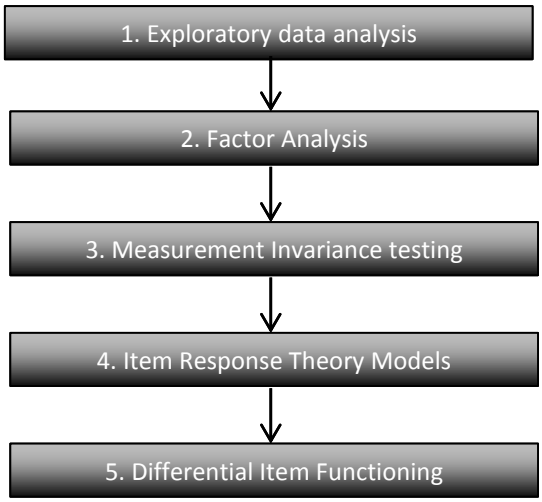
Analysis

The analysis consisted of a multi-step Item Response Theory (IRT) based structural equation modeling (SEM) approach. The analysis allowed for the examination of both the latent structure and the individual items of the HSCL across the study settings. The steps (Figure 2) involved: 1) exploratory data analysis including examination of demographics, distributions of item responses, and summary statistics on average scores for each scale in each setting; 2) Principal Components Analysis (PCA), Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) to examine the dimensionality and underlying factor structure of the data;

3) testing of measurement invariance comparing each setting to all other settings in the overall dataset to determine whether (a) the HSCL had the same underlying factor structure across settings, (b) factor loadings were the same across settings, and (c) item intercepts were the same across settings; 4) fitting of an IRT model to examine the performance of each item across settings and within each individual setting; and 5) an investigation into differential item functioning (DIF) by setting to see if responses to items were systematically different depending on which setting respondents were from. Analyses were done using a variety of statistical software including STATA v. 12 (StataCorp, 2011), Mplus v7.3 (Muthén & Muthén, 1998-2012) and IRTPro (Cai, Du Toit, & Thissen, 2011).

Figure 5.2

Multi-step Item Response Theory (IRT) analysis process



Step 1: Exploratory analysis

Exploratory data analysis was done to examine basic descriptive statistics, distribution of item responses, and average scores across all items on the HSCL-15 across settings. Internal

consistency reliabilities were estimated for the HSCL-15 in each setting separately and across all settings combined.

Step 2: Factor analysis

Factor analyses were done using the full dataset (i.e. combined data from all cultural settings). Principal Components Analysis (PCA), Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were used to assess the factor structure and dimensionality of the data. All three factor analytic methods were used complimentary: PCA examines the underlying correlation matrix and can be used as a preliminary step to inform the number of potential factors underlying the data; EFA can then be used to examine the strength of association of the items to each factor (number of factors suggested by PCA); and CFA allows for testing of the model fit to the data. Using all three of these strategies allowed for first exploring the dimensionality of the data (using PCA and EFA) and then testing the proposed dimensional model to see if it accurately fit the data.

A PCA using a polychoric correlation matrix to account for categorical response categories and a corresponding scree plot were generated to examine underlying components of the data. For the EFA and CFA analysis, the overall sample was randomly split into two groups of about equal size: a development sample and a validation sample. The EFA was conducted using the data from the development sample and the CFA was done using data from the validation sample. Both the EFA and CFA were done using a mean and variance-adjusted weighted least squares estimator (WLSMV) in Mplus 7.3 (Muthén & Muthén, 1998-2012). Geomin rotated standardized factor loadings were examined to see the degree to which individual items loaded on a single factor. Absolute fit of the confirmatory models was evaluated using global fit indices, including the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). RMSEA values lower than 0.06 and TLI/CFI values above 0.95 are indicative of good model fit (Hu & Bentler, 1998).

Step 3: Measurement invariance

To test for measurement invariance across settings, the data were analyzed comparing each setting to all other data in the dataset (e.g. Colombia vs. all others). Configural, metric, and scalar invariance were tested across settings. Configural invariance tests if the same set of factors is present and indicates if the factor structure of the measure (i.e. HSCL-15) is similar across settings. Metric invariance tests if factor loadings are the same across settings and indicates whether the items are correlated with similar magnitudes to the underlying latent trait (i.e. depression) across settings. Scalar invariance is more restrictive than metric invariance and tests if item thresholds and factor loadings are the same, and reflects whether there are systematic differences in the way individuals from different settings respond to the items. Fit of each invariance model was evaluated using global fit indices as described above (Hu & Bentler, 1998). To compare unconstrained models (allowing parameters to vary by setting) and constrained models, we performed chi-squared difference tests ($\Delta\chi 2$) to determine whether the difference between model fit was significant. Non-significant ($p < 0.05$) chi-squared difference tests indicate that the more constrained model does not worsen model fit and may be accepted as an adequate model to explain the data..

Step 4: Item response theory model

IRT-based item discrimination (a) and item location (b) parameters, were estimated for each item on the HSCL-15 using a graded response model (Samejima, 1997). Parameters were estimated using the whole dataset and then for each setting separately. Item discrimination parameters (a) are analogous to factor loadings for each item and indicate how strongly an item is correlated to the underlying latent trait. Item discrimination parameters (a) can also be interpreted as how well the item discriminates between people with different levels of the latent trait.

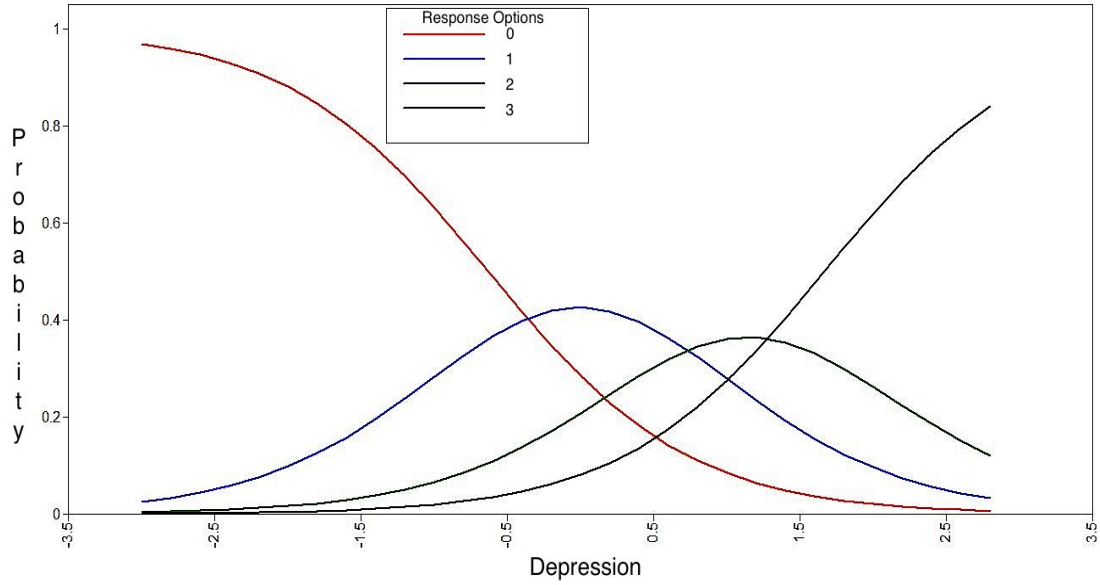
Generally, item discrimination values of 0.01-0.34 are considered very low; 0.35-0.64 low; 0.65-1.34 moderate; 1.35-1.69 high; and 1.70 and above, very high (Baker, 2001).

Item location parameters (b) are defined as the amount or level of the underlying latent trait (θ) where the probability of endorsing the item (or endorsing the particular response category) is 0.50. As each item on the HSCL-15 is assessed based on frequency of experiencing it (ranging from *0 = not at all* to *3 = extremely*), three item location parameters (b_1, b_2, b_3) were estimated. The first location parameter (b_1) represents the level of the underlying latent trait where the probability of endorsing the item with a "0" instead of a "1," "2," or "3" is 0.50. The second location parameter (b_2) represents level of the underlying latent trait where the probability of endorsing the item with a "0" or "1" instead of a "2" or "3" is 0.50. The third location parameter (b_3) is for the response of a "0," "1," or "2" instead of a "3."

Item discrimination parameters (a) and item location parameters (b_1, b_2, b_3) can be depicted graphically using item characteristic curves (ICCs). In a graded response model, these ICCs are called category response curves (CRC) and represent the item parameters for each possible response option for an item. The discrimination parameter corresponds to the steepness of the curves. The location parameters do not directly correspond to the CRCs, but are based on calculating the difference between each curve (Figure 5.3). Discrimination parameters, location parameters, CRCs and item-level information were used to examine the performance of the HSCL-15 items across and within each of the study settings. (see Appendix D for ICCs and information curves for all items in each setting). A test information curve (TIC) was produced to identify the range of latent trait values (i.e. severity levels of depression) for which the scale is most reliable (Hambleton et al., 1991). All IRT models were done using IRTPRO (Cai et al., 2011).

Figure 5.3

Example of an Item Characteristic Curve (ICC) for a graded response model with four Category Response Curves (CRCs)



Step 5: Differential item functioning

Differential Item Functioning (DIF) or item bias occurs when respondents with the same level of latent trait (i.e. same severity of depression) have different probabilities of endorsing an item based on some other characteristic (i.e. ethnicity, gender, age). DIF by setting was evaluated for each of the items in the HSCL-15 by comparing data from one setting (comparison group) to data from all other settings in the dataset (reference group).

Comparability of groups

DIF was evaluated for a total of eight comparison groups (representing each study setting). There were a total of eight different reference groups with each group having one setting removed at a time (i.e. *group 1* included data from all settings but without the Colombia sample;

group 2 included data from all settings but without the Indonesian sample, etc.). For the DIF analysis, it was important to compare the means and standard deviations of scores on the HSCL-15 for the different references groups. If the resulting means and standard deviations of the reference groups were relatively similar, then estimates between comparison groups could be evaluated. For example, if the means and standard deviations for each reference group were similar, than an item that shows high magnitude DIF in Colombia can then be compared to the magnitude of DIF for that same item in Indonesia.

DIF Detection

There are two types of DIF that can be estimated: *non-uniform DIF*, or DIF in the discrimination parameters; and *uniform DIF*, or DIF in the location parameters. *Non-uniform DIF* is analogous to effect modification and represents an interaction between the level of the latent trait (i.e. depression), group membership (i.e. study setting) and the item responses. *Uniform DIF* is analogous to confounding or when the differences in responses to items can be found at all levels of the latent trait (Crane, Belle, & Larson, 2004).

Non-Uniform and *uniform DIF* were identified using MIMIC (multiple indicator, multiple causes) models with a WLSMV estimator. MIMIC models model the effect of a covariate on the latent variable, the factor loading and the threshold of the each item simultaneously (Woods, 2009b; Woods & Grimm, 2011). All MIMIC models were run using Mplus v7.3 (Muthén & Muthén, 1998-2012) with a Bonferroni correction to adjust the level of significance to account for multiple comparisons.

Evaluation of the impact of DIF

To investigate the salience of any non-uniform and/or uniform DIF detected, differences in latent mean scores for depression were examined using models that accounted for both types of DIF compared to models that did not account for the DIF. If the difference in latent mean score

for depression between the two models was statistically significant, then the DIF was considered to have a salient impact on scale level scores.

To determine salience of DIF, anchor items, or items free of non-uniform and uniform DIF across settings, had to be identified. Anchor items were identified from the MIMIC models. If no anchor items could be identified, then a procedure laid out by Woods et al. (2009a) was used. The Woods method involves comparing nested models, by testing a fully constrained model against a model that allows item loadings and thresholds to vary by setting. A logliklihood ratio (LR) statistic is generated which is equal to -2 times the difference between the optimized logliklihoods of the two models divided by the number of free parameters for each item. The LR statistics for each item are then ranked from smallest to largest and the items with the smallest LR statistics are then designated to be anchor items (Woods, 2009a).

Results

Exploratory analysis

The data used in this analysis are described in Table 1. Combining all the data resulted in a dataset with $N = 4732$ individuals. Over half of the participants were women (62.1%), most were between the ages of either 25-44 (46.5%) or 45-66 (28.6%) and a little less than half reported being married (42.8%) at the time of the studies (Table 2). In terms of item level response distributions across all datasets, most items were skewed with the majority of people indicating either 0 “not at all” or 1 “A little bit” (Table 3). Mean scores on the HSCL-15 (possible range 0-3) ranged from $\mu = 0.61$ in Rwanda to $\mu = 1.47$ in Dohuk. Across datasets the internal consistency reliability (α) for the HSCL was good ($\alpha = 0.87$), ranging from $\alpha = 0.79$ in Dohuk to $\alpha = 0.93$ in Southern Iraq.

Table 5.2

Sample Characteristics (N=4732)

Gender *N* (%)	
Male	1778 (37.6)
Female	2939 (62.1)
Missing	15 (0.3)
Age *N* (%)	
14-16	5 (0.1)
16-24	834 (17.6)
25-44	2200 (46.5)
45-66	1353 (28.6)
67-79	247 (5.2)
80+	75 (1.6)
Missing	18 (0.4)
Marital status *N* (%)	
Married	2024 (42.8)
Other	1723 (36.4)
Missing	985 (20.8)
Education *N* (%)	
None	959 (20.3)
Primary	1899 (40.1)
Secondary	701 (14.8)
More than secondary	513 (10.8)
Missing	660 (14.0)*

*Indonesian sample did not report education level

Table 5.3

Number of respondents reporting each response (%) for each item on the HSCL-15 (N=4732)

Please describe how much the symptoms bothered you or distressed over the last period of time[a]	Not at all 0	A little 1	Quite a bit 2	Extremely 3	missing
1. Hopeless	2165 (45.8)	905 (19.1)	946 (20.0)	705 (14.9)	11 (0.2)
2. Crying	1892 (40.0)	953 (20.1)	1065 (22.5)	815 (17.2)	7 (0.2)
3. Sad	1018 (21.5)	1147 (24.2)	1439 (30.4)	1123 (23.7)	5 (0.1)
4. Lonely	1910 (40.4)	949 (20.1)	1099 (23.2)	757 (16.0)	17 (0.4)
5. Sexual interest[b]	1984 (41.9)	781 (16.5)	611 (12.9)	668 (14.1)	688 (14.5)
6. Lack of interest	2273 (48.0)	1098 (23.2)	971 (20.5)	372 (7.9)	18 (0.4)
7. Low energy	1204 (25.4)	1281 (27.1)	1422 (30.1)	822 (17.4)	3 (0.1)
8. Poor appetite	1790 (37.8)	1262 (26.7)	1108 (23.4)	567 (12.0)	5 (0.1)
9. Problems with sleep	1646 (34.8)	860 (18.2)	1285 (27.2)	935 (19.8)	6 (0.1)
10. Thoughts of death	3701 (78.2)	565 (11.9)	322 (6.8)	141 (3.0)	3 (0.1)
11. Trapped	2272 (48.0)	1093 (23.1)	901 (19.0)	457 (9.7)	9 (0.2)
12. Worry[b]	958 (20.3)	888 (18.8)	1285 (27.2)	1000 (21.1)	601 (12.7)
13. Blaming self	2257 (47.7)	1118 (23.6)	957 (20.2)	389 (8.2)	11 (0.2)
14. Effort	1382 (29.2)	1028 (21.7)	1298 (27.4)	1006 (21.3)	18 (0.4)
15. Worthlessness	2830 (59.8)	852 (18.0)	679 (14.4)	352 (7.4)	19 (0.4)

[a] Time frame of the question varied by study and ranged from 1-4 weeks.

[b] not asked in the Indonesian data

Factor analysis

Principal Components Analysis (PCA) and the associated scree plot indicated one predominant factor with an eigenvalue of 6.6. The next highest eigenvalue was 1.1. The

development sample (n = 2407) was used for the EFA and the validation sample (n = 2325) for the CFA. Results from the EFA and CFA supported a unidimensional construct. Factor loadings from the 1-factor EFA ranged from 0.50 for the item "Loss of sexual interest/pleasure" to 0.79 for the item "Feeling sad." The CFA of a 1-factor model yielded excellent model fit indices (RMSEA = 0.05; CFI = 0.95; TLI = 0.94)

Measurement invariance

Configural measurement invariance was largely supported in all settings, demonstrating that a 1-factor structure fits the data in all settings. RMSEA values for all configural models ranged from 0.064 in Dohuk to 0.075 in Colombia, Thailand, and Uganda. CFI and TLI values for configural invariance models were all above 0.90. Model fit indices indicated that metric and scalar invariance showed less adequate fit. Metric model fit indices ranged from RMSEA values of 0.06 to 0.09; CFI values of 0.915 to 0.959 and TLI values of 0.908 to 0.956. The metric and scalar models comparing Thailand to all other settings had the worse model fit indicating questionable measurement invariance for this comparison: RMSEA = 0.087, CFI = 0.915 and TLI = 0.908 for the metric model and RMSEA = 0.089, CFI = 0.897 and TLI = 0.903 for the scalar model. Across all settings RMSEAs were consistently either just below or over the cut off for good fit but CFIs and TLIs indicated good model fit (with the exception of the scalar models in Thailand) However, all chi-squared difference tests were significant indicating that constrained models (i.e. metric and scalar models) did not fit the data better than unconstrained models. (Tables 4). These results show that the same underlying unidimensional model is consistent across settings, but factor loadings and thresholds vary by setting.

Table 5.4

Model fit statistics for configural, metric, and scalar invariance models comparing each setting to all other settings (n = 4732)

Model	X^2	df	P value	RMSEA[a]	CFI[a]	TLI[a]
Colombia v. All others						
Configural	2589.745	180	0.000	0.075	0.946	0.937
Metric	3466.357	194	0.000	0.084	0.927	0.921
Scalar	4552.835	223	0.000	0.091	0.903	0.909
Dohuk v. All others						
Configural	1949.568	180	0.000	0.064	0.959	0.952
Metric	1940.615	194	0.000	0.062	0.959	0.956
Scalar	2635.480	223	0.000	0.068	0.944	0.947
Kudistan v. All others						
Configural	2435.714	180	0.000	0.073	0.946	0.937
Metric	2534.913	194	0.000	0.071	0.944	0.939
Scalar	3195.917	223	0.000	0.075	0.928	0.932
Indonesia v. All others						
Configural	1436.990	130	0.000	0.065	0.965	0.958
Metric	1961.470	142	0.000	0.074	0.951	0.946
Scalar	3412.700	167	0.000	0.091	0.912	0.918
Iraq v. All others						
Configural (A)	2173.880	180	0.000	0.068	0.949	0.941
Metric (B)	2245.121	194	0.000	0.067	0.948	0.944
Scalar (C)	2132.732	223	0.000	0.060	0.952	0.954
Rwanda v. All others						
Configural	2236.714	180	0.000	0.069	0.948	0.940

Metric	2381.779	194	0.000	0.069	0.945	0.941
Scalar	2242.387	223	0.000	0.062	0.949	0.952
Thailand v. All others						
Configural	2571.792	180	0.000	0.075	0.941	0.931
Metric	3637.814	194	0.000	0.087	0.915	0.908
Scalar	4394.687	223	0.000	0.089	0.897	0.903
Uganda v. All others						
Configural	2605.328	180	0.000	0.075	0.945	0.935
Metric	3136.620	194	0.000	0.080	0.933	0.927
Scalar	2982.130	223	0.000	0.072	0.937	0.941

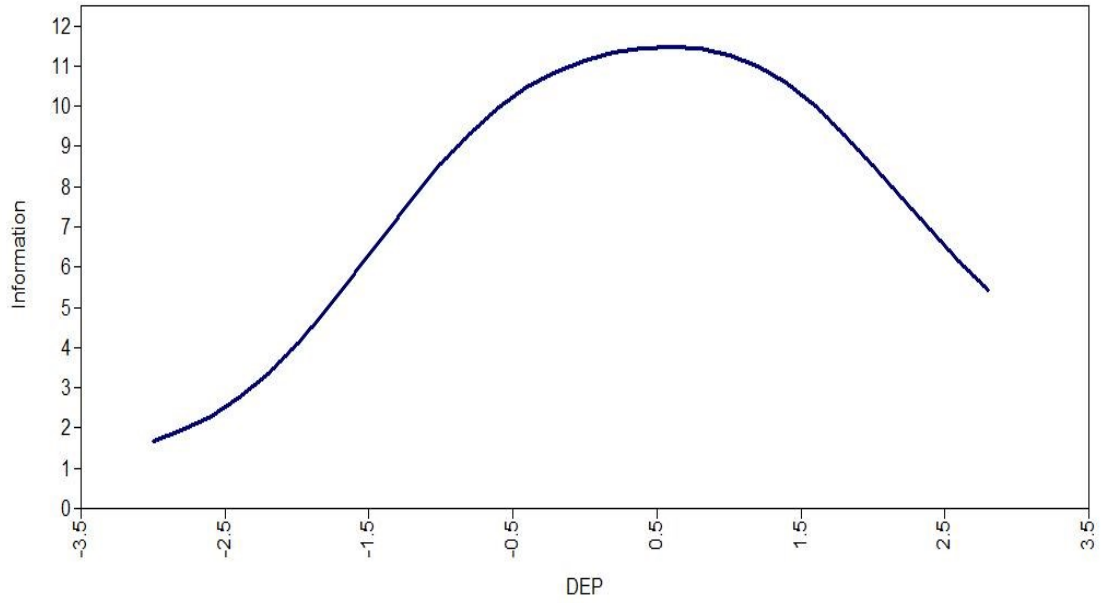
[a]RMSEA = root mean square error of approximation; CFI = the comparative fit index; TLI = Tucker-Lewis index. Good model fit indicated by RMSEA values lower than 0.05 and TLI/CFI values above 0.90

Item response theory model

The overall IRT analysis of the HSCL-15 in the combined dataset (*N = 4732*) indicated most items performed well. Item discrimination parameters ranged from a = 0.97 ("lack of interest or pleasure in sex") to a = 2.09 ("feel sad") (Table 5). Item location parameters (for a response of 0 compared to a response of 1, 2, or 3) ranged from b = -1.02 ("feel sad") to b= 1.38 ("thoughts of death/suicide") (Table 6). The HSCL-15 was most reliable for people in the Θ = 0.5 to Θ = 1.0 range indicating that the scale is most reliable for people with depression slightly above the average level of depression across datasets (Figure 4).

Figure 5.4

Test information curve for HSCL-15 in combined sample (N=4732)



In the setting-specific IRT analyses, discrimination parameters ranged from $a = 0.31$ for the item "thoughts of death/suicide" in Thailand to $a = 3.10$ for the item "no interest" in Northern Iraq. The value of the discrimination parameter for the item "thoughts of death/suicide" in Thailand is considered very low, indicating that this item may not be related to depression in this setting. In Northern Iraq, the item "no interest" showed a very high discrimination parameter indicating the item is strongly correlated to depression and can accurately discriminate between people with different levels of depression. The item "lack of interest or pleasure in sex" consistently had low or moderate discrimination parameters across settings; $a = 0.74$ in Rwanda to $a = 1.26$ in Northern Iraq. Other notably low discrimination parameters were observed for the item "feeling everything is an effort" in Colombia ($a = 0.88$) and the items "loss of interest" ($a = 0.74$) and "self-blame" ($a = 0.98$) in Rwanda (Table 5).

Across settings the item "crying" was most commonly endorsed by individuals with low levels of depression (location parameters in overall sample: b_1 = -1.02; b_2 = -0.10; b_3 = 0.94), while the item "thoughts of death or suicide" was most commonly endorsed by people with higher levels of depression (location parameters in overall sample: b_1 = 1.38; b_2 = 2.33; b_3 = 3.58). Dohuk had the lowest item location parameter for the item "crying" (b = -3.67 for a response of 0 compared to 1, 2 or 3) meaning that individuals in Dohuk with low levels of depression were very likely to endorse this item. In Thailand, the item "thoughts of death/suicide" was relatively "difficult" (b = 2.74 for a response of 0 compared to 1, 2 or 3) meaning that individuals had to have high levels of depression to endorse this item (Table 5.6).

Table 5.5

Item discrimination parameters (a) and their standard errors: overall and by setting (N = 4732)

	Overall	Colombia	Indonesia[a]	Dohuk	Rwanda	Thailand	S. Iraq	Uganda	N.Iraq
Hopeless (d01)	1.75 (0.05)	1.79 (0.58)	1.92 (0.26)	1.94 (0.39)	1.72 (0.26)	1.93 (0.19)	2.74 (0.44)*	1.44 (0.18)	2.19 (0.31)
Crying (d02)	1.50 (0.05)	1.56 (0.49)	1.26 (0.17)	0.69 (0.28)*	1.47 (0.25)	1.38 (0.15)	1.43 (0.30)	1.34 (0.17)	1.33 (0.21)*
Sad (d03)	2.09 (0.06)	2.26 (0.66)	1.71 (0.22)	2.29 (0.57)*	2.83 (0.39)	2.79 (0.29)*	2.98 (0.46)	1.86 (0.23)	2.68 (0.30)
Lonely (d04)	1.89 (0.06)	1.98 (0.57)	2.22 (0.29)	2.01 (0.46)	2.24 (0.31)	2.14 (0.21)	2.81 (0.42)	1.58 (0.19)	1.76 (0.24)
Lost sex (d05)	0.97 (0.04)	1.01 (0.28)		0.91 (0.33)	0.74 (0.16)	1.17 (0.15)*	1.00 (0.25)	0.85 (0.13)	1.26 (0.16)
No interest (d06)	1.44 (0.05)	0.98 (0.27)*	1.69 (0.21)	1.76 (0.46)	0.76 (0.18)	1.70 (0.17)	2.70 (0.42)*	1.67 (0.19)	3.10 (0.33)*
Low energy (d07)	1.28 (0.04)	1.32 (0.35)*	1.41 (0.20)	1.70 (0.46)	1.55 (0.21)	2.01 (0.20)*	2.15 (0.36)	1.36 (0.16)	1.46 (0.20)
Appetite (d08)	1.05 (0.04)	1.09 (0.32)	1.23 (0.18)	1.41 (0.53)	1.60 (0.22)	1.05 (0.13)	1.44 (0.30)	1.11 (0.16)	1.33 (0.19)
Sleep (d09)	1.29 (0.04)	1.30 (0.39)*	1.24 (0.19)	1.67 (0.37)	1.66 (0.23)	2.23 (0.22)*	1.91 (0.35)	1.46 (0.18)	1.17 (0.19)*
Suicide (d10)	1.12 (0.04)	1.41 (0.41)	2.49 (0.47)	1.98 (0.32)	2.27 (0.52)	0.31 (0.10)*	1.59 (0.35)	1.03 (0.22)	1.60 (0.24)
Trapped (d11)	1.14 (0.04)	1.46 (0.43)	1.52 (0.21)	2.08 (0.33)	2.68 (0.39)*	2.01 (0.19)*	2.49 (0.38)*	1.26 (0.16)	1.83 (0.27)
Worry (d12)	1.57 (0.05)	1.21 (0.38)*		1.62 (0.53)*	1.35 (0.21)	2.83 (0.26)*	1.70 (0.29)	1.52 (0.18)	1.52 (0.22)*
Self-blame (d13)	1.22 (0.04)	1.10 (0.33)*	1.29 (0.19)	1.48 (0.33)	0.98 (0.19)	1.66 (0.16)*	1.77 (0.32)	0.63 (0.12)	1.80 (0.25)
Effort (d14)	1.47 (0.04)	0.88 (0.27)*	1.75 (0.24)	1.67 (0.42)	2.08 (0.27)	2.03 (0.19)*	2.48 (0.42)	1.88 (0.20)	1.90 (0.24)

Worthlessness (d15)	1.47 (0.05)	1.48 (0.43)*	2.84 (0.29)	1.28 (0.31)	2.43 (0.32)	1.63 (0.17)	2.65 (0.40)*	1.58 (0.19)	1.89 (0.26)

[a] Items left blank were not included in the Indonesia dataset

*Identifies statistically significant non-uniform DIF ($p < 0.001$) when comparing the setting in each column to all other settings in the combined dataset using MIMIC models

Table 5.6

Item location parameters (b_1, b_2, b_3) and their standard errors: overall and by setting (N = 4732)[a]

Threshold	Overall	Colombia	Indonesia[b]	Dohuk	Rwanda	Thailand	S. Iraq	Uganda	N.Iraq
D1 Hopeless, b_1	-0.14 (0.02)	-0.79 (0.39)	0.43 (0.09)*	-1.30 (0.31)	-0.53 (0.12)	-0.90 (0.11)*	-0.62 (0.12)	-0.39 (0.11)	-0.76 (0.11)
D1 Hopeless, b_2	0.54 (0.03)	-0.15 (0.19)	0.79 (0.11)	0.07 (0.14)	0.24 (0.14)	-0.20 (0.08)*	-0.09 (0.10)	0.14 (0.11)	-0.03 (0.07)
D1 Hopeless, b_3	1.45 (0.04)	0.59 (0.08)*	1.32 (0.17)	1.18 (0.16)	1.56 (0.31)	0.85 (0.10)	0.72 (0.12)	0.85 (0.16)	1.10 (0.16)*
D2 Crying, b_1	-0.39 (0.03)	-1.10 (0.48)	-0.56 (0.11)	-3.67 (1.42)	0.07 (0.14)	-0.94 (0.12)	-0.85 (0.17)	-0.39 (0.10)	-1.40 (0.20)
D2 Crying, b_2	0.38 (0.03)	-0.40 (0.27)*	0.11 (0.10)	-0.57 (0.26)	0.86 (0.23)	0.13 (0.08)	0.23 (0.16)	0.13 (0.11)	-0.36 (0.09)
D2 Crying, b_3	1.42 (0.04)	0.46 (0.06)*	0.88 (0.16)	2.55 (1.06)	2.47 (0.54)	1.70 (0.21)	0.97 (0.25)	0.87 (0.16)	1.35 (0.24)
D3 Sad, b_1	-1.02 (0.03)	-1.62 (0.63)	-1.68 (0.21)*	-1.10 (0.33)	-1.41 (0.16)	-1.37 (0.14)	-1.02 (0.14)	-1.34 (0.16)	-1.38 (0.16)
D3 Sad, b_2	-0.10 (0.02)	-0.71 (0.36)	-0.95 (0.14)*	-0.03 (0.17)	-0.45 (0.09)	-0.52 (0.08)	0.09 (0.09)	-0.55 (0.10)	-0.59 (0.09)
D3 Sad, b_3	0.94 (0.03)	0.07 (0.14)*	0.02 (0.09)*	1.23 (0.17)	0.86 (0.15)	0.74 (0.08)	0.84 (0.13)	0.20 (0.09)	0.67 (0.09)*
D4 Lonely, b_1	-0.31 (0.02)	-0.95 (0.44)	-0.27 (0.08)	-1.10 (0.33)	-0.74 (0.11)	-0.65 (0.10)	-0.89 (0.13)	-1.00 (0.14)	-0.94 (0.13)
D4 Lonely, b_2	0.38 (0.02)	-0.31 (0.25)	0.12 (0.07)	-0.03 (0.17)	-0.16 (0.10)	-0.00 (0.08)	0.21 (0.10)	-0.32 (0.09)	-0.16 (0.08)
D4 Lonely, b_3	1.34 (0.04)	0.43 (0.06)*	0.70 (0.10)	1.23 (0.17)	1.19 (0.22)	0.91 (0.11)	1.16 (0.17)	0.70 (0.12)	1.29 (0.19)*
D5 Lost sex, b_1	-0.02 (0.04)	-0.51 (0.32)	--	-1.01 (0.31)	-0.82 (0.17)	0.15 (0.11)*	-1.02 (0.26)	-0.59 (0.14)	-0.75 (0.12)
D5 Lost sex, b_2	0.97 (0.05)	0.42 (0.11)	--	0.55 (0.30)	0.42 (0.3)	1.08 (0.19)*	0.55 (0.28)	0.15 (0.16)	0.06 (0.09)

D5 Lost sex, b_3	1.97 (0.08)	0.98 (0.15)*	--	2.41 (0.90)	2.04 (0.61)	2.59 (0.38)	1.72 (0.50)	1.11 (0.26)	1.22 (0.17)
D6 No interest, b_1,	-0.08 (0.03)	-0.41 (0.28)*	0.31 (0.09)	-1.63 (0.28)	0.49 (0.34)	-0.89 (0.11)*	-1.11 (0.16)	-0.22 (0.10)	-1.04 (0.13)*
D6 No interest, b_2	0.85 (0.03)	0.78 (0.12)*	0.88 (0.14)	-0.02 (0.16)	2.57 (0.80)	0.04 (0.08)*	0.09 (0.10)	0.49 (0.11)	-0.24 (0.07)*
D6 No interest, b_3	2.19 (0.06)	2.01 (0.43)	1.93 (0.26)	1.73 (-0.42)	5.54 (1.64)	1.74 (0.20)	1.13 (0.18)	1.53 (0.20)	0.83 (0.11)
D7 Low energy, b_1,	-1.07 (0.04)	-1.43 (0.56)	-2.02 (0.26)*	-2.17 (0.45)	-1.83 (0.21)	-0.83 (0.11)*	-1.53 (0.21)	-1.24 (0.16)	-2.11 (0.27)
D7 Low energy, b_2	0.14 (0.03)	-0.24 (0.23)*	-1.13 (0.16)*	-0.26 (0.13)	-0.72 (0.13)	0.06 (0.08)	-0.08 (0.11)	-0.53 (0.11)	-0.51 (0.10)
D7 Low energy, b_3	1.56 (0.05)	0.98 (0.15)	-0.14 (0.10)*	1.29 (0.38)	0.89 (0.21)	1.35 (0.15)	1.01 (0.19)	0.56 (0.13)	1.40 (0.21)
D08 Appetite, b_1,	-0.58 (0.04)	-0.84 (0.41)*	-1.84 (0.24)*	-1.73 (0.47)	-1.08 (0.15)	-1.24 (0.15)	-1.31 (0.22)	0.23 (0.13)*	-1.09 (0.15)
D08 Appetite, b_2	0.68 (0.04)	0.22 (0.13)*	-1.09 (0.16)*	0.21 (0.20)	-0.03 (0.13)	0.62 (0.14)	0.14 (0.15)	0.95 (0.19)	0.09 (0.09)
D08 Appetite, b_3	2.22 (0.08)	1.39 (0.28)	0.42 (0.13)*	1.54 (0.58)	1.43 (0.28)	2.84 (0.39)	1.64 (0.37)	1.89 (0.31)	1.83 (0.27)
D9 Sleep, b_1,	-0.63 (0.03)	-0.87 (0.41)*	-2.01 (0.26)*	-1.63 (0.29)	-1.17 (0.15)	-0.80 (0.10)	-1.55 (0.23)	-0.48 (0.11)*	-1.57 (0.22)
D9 Sleep, b_2	0.14 (0.03)	-0.22 (0.22)*	-1.30 (0.18)*	-0.18 (0.15)	-0.33 (0.13)	-0.42 (0.08)	-0.30 (0.12)	0.05 (0.10)	-0.44 (0.10)
D9 Sleep, b_3	1.40 (0.05)	0.73 (0.11)	-0.02 (0.10)*	1.07 (0.32)	1.43 (0.28)	0.86 (0.10)	0.77 (0.18)	0.73 (0.14)	1.24 (0.22)
D10 Suicide, b_1,	1.38 (0.05)	0.87 (0.14)*	1.46 (0.21)	-0.68 (0.20)*	0.85 (0.23)	2.74 (1.08)	0.83 (0.20)	2.36 (0.53)	0.37 (0.10)
D10 Suicide, b_2	2.33 (0.09)	1.45 (0.30)	1.75 (0.25)	0.39 (0.18)*	1.12 (0.29)	9.12 (3.06)	1.73 (0.38)	2.79 (0.61)	1.12 (0.18)
D10 Suicide, b_3	3.58 (0.15)	2.22 (0.51)	2.14 (0.31)	1.80 (0.30)	2.30 (0.64)	13.38 (4.45)	2.57 (0.61)	3.28 (0.72)	2.32 (0.34)
D11 Trapped, b_1,	-0.11 (0.03)	0.16 (0.12)*	-0.23 (0.09)	-0.87 (0.20)	-0.44 (0.10)	-1.62 (0.16)*	-0.60 (0.12)	-0.94 (0.13)*	-0.52 (0.10)
D11 Trapped, b_2	0.96 (0.04)	0.81 (0.13)*	0.45 (0.11)	0.29 (0.18)	0.07 (0.09)	-0.32 (0.08)	0.28 (0.11)	-0.23 (0.10)*	0.29 (0.09)

D11 Trapped, b_3	2.35 (0.08)	1.60 (0.36)*	1.06 (0.17)*	1.41 (0.25)	1.27 (0.23)	1.19 (0.13)	1.20 (0.19)	0.72 (0.15)	1.61 (0.23)
D12 Worry, b_1,	-1.05 (0.04)	-2.30 (0.83)*	--	-1.53 (0.33)	-0.87 (0.14)	-1.22 (0.13)	-1.56 (0.21)	-1.64 (0.19)	-1.59 (0.21)
D12 Worry, b_2	-0.14 (0.03)	-1.21 (0.50)*	--	-0.19 (0.15)	0.19 (0.16)	-0.59 (0.09)	-0.36 (0.12)	-0.66 (0.11)	-0.66 (0.11)
D12 Worry, b_3	1.07 (0.04)	-0.06 (0.16)*	--	1.86 (0.47)	2.29 (0.48)	0.54 (0.07)	0.69 (0.17)	0.42 (0.11)	1.04 (0.18)*
D13 Self-blame, b_1,	-0.11 (0.03)	-0.50 (0.30)	0.44 (0.12)	-2.04 (0.37)	-0.04 (0.20)	-1.04 (0.12)*	-1.70 (0.24)	0.95 (0.31)*	-1.00 (0.13)*
D13 Self-blame, b_2	0.93 (0.04)	0.45 (0.09)	1.13 (0.20)	0.07 (0.19)	1.76 (0.49)	-0.04 (0.08)*	-0.25 (0.12)	2.62 (0.61)	-0.03 (0.08)
D13 Self-blame, b_3	2.40 (0.08)	1.70 (0.40)	2.33 (0.37)	1.86 (0.47)	3.27 (0.81)	1.49 (0.17)	0.79 (0.19)	4.11 (0.91)	1.49 (0.22)
D14 Effort, b_1,	-0.81 (0.03)	-2.69 (0.94)*	-1.33 (0.17)	-1.71 (0.31)	-1.23 (0.15)	-0.62 (0.09)*	-0.85 (0.14)	-1.14 (0.14)	-1.25 (0.16)
D14 Effort, b_2	0.09 (0.03)	-1.34 (0.53)*	-0.72 (0.11)*	-0.17 (0.16)	-0.36 (0.11)	0.01 (0.08)	0.14 (0.11)	-0.47 (0.10)	-0.24 (0.08)
D14 Effort, b_3	1.23 (0.04)	0.14 (0.13)*	0.12 (0.09)*	1.45 (0.40)	1.14 (0.22)	1.29 (0.14)	1.17 (0.20)	0.39 (0.10)	1.20 (0.18)*
D15 Worthless, b_1,	0.38 (0.03)	0.32 (0.08)*	0.25 (0.07)	-1.18 (0.24)	-0.43 (0.11)	-0.20 (0.08)	-0.57 (0.11)	-0.43 (0.10)	-0.48 (0.09)
D15 Worthless, b_2	1.16 (0.04)	0.96 (0.17)*	0.68 (0.09)	0.32 (0.24)	0.13 (0.11)	0.63 (0.11)	0.04 (0.10)	0.25 (0.10)	0.29 (0.09)
D15 Worthless, b_3	2.21 (0.07)	1.64 (0.37)	1.19 (0.14)	1.85 (0.52)	1.23 (0.22)	1.85 (0.22)	0.79 (0.15)	1.12 (0.16)	1.59 (0.22)

[a] *b_1 = difficulty parameter for an item-response of 0 instead of 1,2, or 3; b_2 = difficulty parameter for an item-response of 0 or 1 instead of 2 or 3; b_3 = difficulty parameter for an item-response of a 0, 1, or 2 instead of 3.*

[b] Items left blank were not included in the Indonesia dataset

* Identifies statistically significant uniform DIF ($p < 0.001$) comparing the setting in each column to all other settings in the combined dataset using MIMIC models

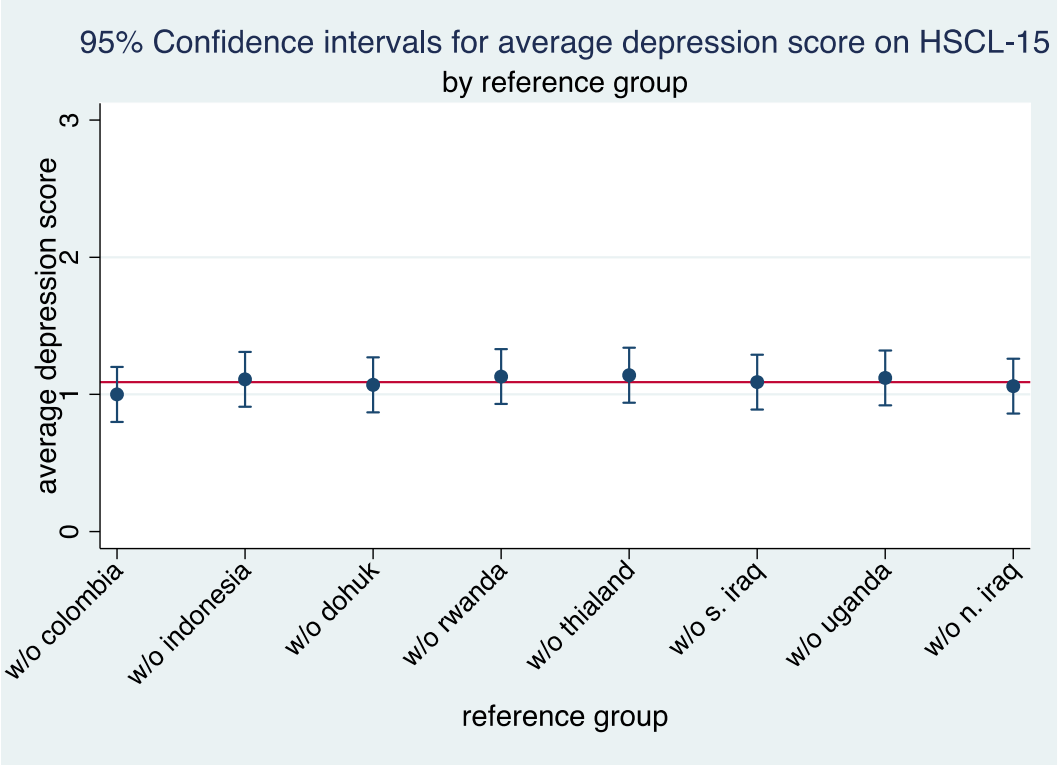
Differential item functioning (DIF)

Comparability of groups

DIF detection was done by comparing scale parameters in one setting (comparison group) to scale parameters in all other settings (reference group) combined (e.g. Colombia vs. All Others). To do this, we examined the means and standard deviations across reference groups. Mean HSCL depression scores did not vary widely with the removal of each dataset. They ranged from 1.00 for the total sample without the Colombia data to 1.14 for the total sample without the Thailand data (Figure 5). These results allow interpretation of DIF between comparison groups.

Figure 5.5

Average scores and 95% confidence intervals on the HSCL-15 for each reference group[a]



[a] The first reference group included data on all settings except Colombia; the second reference group included data on all settings except Indonesia; etc.

*Redline indicates mean score on the HSCL-15 across all reference groups (μ = 1.09)

DIF detection

The presence of non-uniform DIF, or DIF in the discrimination parameters, would suggest that setting modified the relationship between depression and item responses. Non-uniform DIF results are presented in Table 5. Non-uniform DIF is indicated to be present if there is a statistically significant ($p < 0.001$) difference between the discriminations parameters in one

setting (comparison group) compared to discrimination parameters in all other settings combined (reference group). Non-uniform was detected in all settings for at least one item on the HSCL-15 with the exception of Indonesia and Uganda. In Thailand, 9 out of 15 items showed non-uniform DIF. Being part of the study sample from Thailand compared to being part of the study samples in all other settings, modified the relationship of these nine items to underlying levels of depression. For example, the item "low energy" seems closely related to depression and is better at discriminating between levels of depression in Thailand compared to all other settings, but the item "thoughts of death/suicide" seems less related to depression in Thailand than in all other settings. At the individual item level, the items "feel lonely" and "changes in appetite" were free of non-uniform DIF across all setting comparisons. It appears that these two items are similarly related to depression regardless of the setting.

The presence of uniform DIF, or DIF in the location parameters, would suggest that study setting confounds the relationship between depression and item responses. Uniform DIF results are presented in Table 6. Uniform DIF is indicated to be present if there is a statistically significant ($p < 0.001$) difference between the location parameters in one setting (comparison group) compared to the location parameters in all other settings combined (reference group). In Colombia, every item except "blaming oneself" showed uniform DIF indicating that participants in Colombia endorsed the items differently than participants in the other settings, despite the same levels of underlying depression. In Indonesia uniform DIF was present for 7 of the items comparing individuals in Indonesia to individuals in all other settings. All items in Rwanda and Southern Iraq were free of uniform DIF meaning that individuals in these settings with the same severity of underlying depression endorse the items similarly compared to people in all other settings.

Impact of DIF

At least one anchor item was identified using MIMIC models in all setting comparisons, with the exception of Colombia vs. all other settings. For the Colombia vs. all other settings comparisons the method by Woods et al. (2009a), as described previously, was used to identify the three best items to use as anchor items for the investigation of the impact of DIF.

Table 8 presents the impact of DIF on latent mean scores of depression. The most salient impact of item-level DIF on aggregate latent mean scores of depression was observed for Indonesia. Without accounting for DIF, people in Indonesia on average had 0.26 units less depression compared to people in all other settings. Once item level DIF was accounted for, participants in Indonesia on average had 0.88 units less depression compared to people in other settings. This difference of 0.62 points, suggests that by not accounting for DIF, average scores of depression in the Indonesian sample are being overestimated. Salient impact of item-level DIF on latent mean values of depression was observed for Indonesia and Southern Iraq.. There was no impact of item-level DIF on latent mean values of depression in Columbia, Dohuk, Rwanda, Thailand, Uganda and Northern Iraq.

Table 5.7.

Difference in latent meant scores of depression by setting accounting for and not accounting for DIF

	Difference in factor means β (*SE*)
Colombia	
Accounting for DIF	0.69 (0.04)
Not accounting for DIF	0.58 (0.03)
Difference	0.11
Indonesia	
Accounting for DIF	-0.88 (0.05)
Not accounting for DIF	-0.26 (0.04)
Difference	0.62*
Dohuk	
Accounting for DIF	0.68 (0.08)
Not accounting for DIF	0.63 (0.07)
Difference	0.05
Rwanda	
Accounting for DIF	-0.98 (0.06)
Not accounting for DIF	-0.99 (0.07)
Difference	0.01
Thailand	
Accounting for DIF	-0.52 (0.05)
Not accounting for DIF	-0.50 (0.05)

Difference	0.02
S. Iraq	
Accounting for DIF	-0.05 (0.11)
Not accounting for DIF	0.38 (0.11)
Difference	0.43*
Uganda	
Accounting for DIF	-0.33 (0.04)
Not accounting for DIF	-0.42 (0.04)
Difference	0.09
N. Iraq	
Accounting for DIF	0.40 (0.06)
Not accounting for DIF	0.35 (0.05)
Difference	0.05

*difference is statistically significant $p < 0.05$

Discussion

This study analyzed data from populations living in eight different study settings within low- and middle-income countries. The aims of this study were to: 1) evaluate the IRT based parameters of the HSCL-15 to examine item performance across all settings and within each setting individually; and 2) to evaluate the HSCL-15 for DIF across settings. The ultimate goal of this analysis was to inform selection of items for a new instrument to measure depression globally. Items that showed high discrimination parameters at a range of severity levels (i.e. location parameters) for depression across settings, and that were demonstrably less biased, would be selected for the new instrument. Although IRT has many benefits for scale development and refinement, few depression measures have been created using IRT, with the notable

exception of the Kessler Psychological Distress Scale (K-10; Kessler et al. 2002) and the PROMIS initiative (Pilkonis et al., 2011) in the United States, and one scale developed for adolescents in Africa (Betancourt, Yang, Bolton, & Normand, 2014). A depression measure made up of items potentially free of bias across study settings, or at least with the biases identified and accounted for during scoring, could provide more valid estimates of clinical and epidemiological disease burden.

The HSCL-15 demonstrated good to excellent internal consistency within and across settings. This is consistent with previous validation studies testing the psychometric properties of this scale in different LMIC (Ertl et al., 2011; Silove et al., 2007). Exploratory and confirmatory factor analyses indicated that a unidimensional model could be assumed across all settings. Results from the invariance testing, fully supported configural invariance. Model fit statistics for the metric and scalar models showed more mixed results. From the invariance testing, it appears that across all settings a 1-factor model is appropriate but there are some differences in the strength of association (factor loadings) and patterns of endorsement (factor thresholds) for items that is at least partially due to study setting.

Overall, most items on the HSCL-15 consistently showed relatively high discrimination parameters indicating a strong relationship of these items to depression regardless of the study setting. The well-performing items include: “feeling hopeless,” “feeling sad,” “feeling low in energy or slowed down,” “problems with sleep,” “feeling trapped,” “worrying too much,” and “feeling worthless.” In terms of location parameters, all of these well-performing items, except “feeling worthless,” were located on the lower end of the latent trait continuum meaning that these items are most informative for measuring lower levels of depression. However, the item “feeling worthless” had a location parameter of $b_1 = 0.38$ (response of 0 compared to 1,2, or 3) indicating this item taps into higher than average levels of depression compared to the other well performing items.

Across all settings the item "lost of sexual interest or pleasure" had relatively low discrimination parameters (a = 0.74 to a = 1.26) meaning this item seems only moderately related to depression across settings. In many non-Western settings, topics related to sex are often not discussed openly. In response to these cultural norms, researchers have modified (Bass et al. 2012) or considered modifying (Kojima et al., 2002) measures, despite the robust evidence that it is strongly related to depression (Fabre & Smith, 2012; Kennedy & Rizvi, 2009; Kitamura, Hirano, Chen, & Hirata, 2004). However, most of the evidence suggesting this strong association comes from high-income settings, while there is little to no literature on the relevance of this symptom to depression in LMIC. The evidence from the current study demonstrated that this item is not strongly related to depression in these settings and could be considered for removal from the scale for future use.

In the overall dataset, several items covered the same area of the latent trait continuum (as evidenced by similar location parameters). These include the items "feeling sad" (b_1 = -0.39, b_2 = 0.38, b_3 = 1.42) and "feeling lonely" (b_1 = -0.31, b_2 = 0.38, b_3 = 1.34) and the items "feeling trapped" (b_1 = -0.11, b_2 = 0.96, b_3 = 2.35) and "blaming oneself" (b_1 = -0.11, b_2 = 0.93, b_3 = 2.40) suggesting these items provide similar information about depression and were redundant. Items that showed redundancy in location parameters could be considered for removal, but further analysis would need to be done to see how removal of such items impacts the reliability of the overall scale.

In the setting specific analyses, the lowest discrimination parameter was observed for the item "thoughts of killing oneself/suicide" (a = 0.31) in the Thailand data. Frequency of item response categories were comparable in all settings, indicating that this low discrimination parameter in Thailand was likely not due to infrequent endorsement of the item. This item may not be a good indicator for depression in this setting. Perhaps, in this context, suicidal ideation is not being driven by depression, but by other factors instead. Recent findings have shown that while depression is a risk factor for suicide ideation in high-income countries, impulse control

disorders are more strongly associated with thoughts of death and suicide in many LMIC (Nock et al., 2008).

Several other items showed low discrimination parameters in some settings including the item "crying a lot" in Dohuk (a = 0.69), the item "blaming oneself" in Uganda (a = 0.63) and the item "feeling as though everything is an effort" in Colombia (a = 0.88). It may be worth dropping these items for future work in these specific settings. It appears that participants' responses to these items are weakly related to depression and these items poorly discriminate between individuals with different severity of depression in these study settings.

The results from the DIF analysis indicate that almost all items showed some form of DIF across setting comparisons. Non-uniform DIF indicates that setting modifies the relationship of the item to the underlying trait of depression. Only the items "loss of sexual interest or pleasure" and "feeling lonely" were completely free of non-uniform DIF, suggesting that setting does not affect the relationship of these items to depression. The findings from the uniform DIF analysis indicated that for many items, setting confounds the relationship of the item response to depression. It may be that what setting one is from affects the symptoms that are indicative of mild or severe depression. For example, as an individual from Indonesia who only has mild depression might indicate experiencing "low energy and fatigue," but this symptom may be more representative of more severe depression for people in all other settings.

The findings from the current study indicating that no items on the HSCL-15 were completely free of non-uniform and/or uniform DIF across settings, is to be expected given the number of settings examined. These findings are consistent with the few other studies that have looked at DIF for depression measures across countries or settings. In the context of cross-national prevalence studies, Nuevo and colleagues (2009) found that 12 out of the 21 items on the Beck Depression Inventory (BDI) showed evidence of DIF across data from five different European countries. Similarly when comparing samples of university students from the U.S. and Turkey, Canel-Cinarbas et al. (2011) found evidence for DIF in a different set of 12 of the 21

items on the BDI. Taken together, the findings from the current study add to the literature suggesting that measurement items for depression functioning differently across settings and any potential DIF should be identified and accounted for in cross-national studies.

While both uniform and/or non-uniform DIF were found for all items, the impact of this DIF on latent mean scores for depression were mixed. There was little impact of item-level DIF on latent mean depression levels in Colombia, Dohuk, Rwanda, Thailand, Uganda and Northern Iraq. The largest impact of item level DIF on latent mean depression scores was for Indonesia, suggesting overestimation of depression levels in this context when DIF was not accounted for.

In Southern Iraq, item level DIF had a significant impact on scale level scores despite only the presence of uniform DIF. This finding could be related to sample size given that data from Southern Iraq constituted the smallest sample in the combined dataset. Unbalanced sample sizes could have made detection of DIF imprecise and risk of a type II error in detecting uniform DIF may have been possible. However, if the estimates are valid, then the DIF in Southern Iraq compared to all other settings, is found at different levels of the latent trait (i.e. people with high levels of depression in Southern Iraq endorse items differently than people with high levels of depression in other settings).

The impact of item-level DIF on aggregate scale scores may be one source of measurement error contributing to heterogeneity of prevalence estimates across countries. Although a recent review by Kessler & Bromet (2014) suggested that measurement factors may not play a significant role in variability in cross-national estimates of depression, this review did not consider differential item functioning as one source of possible measurement error. The evidence from the current study indicates that in some cases, item-level response bias has a significant impact on aggregate estimates of depression. However, the lack of epidemiologic samples in the present study limits the generalizability of the findings related to DIF impact. Future, epidemiologic studies should investigate the potential for item-level DIF to contribute to

heterogeneity in prevalence estimates and if present, account for it when comparing measurement estimates across settings.

Overall, the results from all of the analyses show that most of the items from the HSCL 15-item depression scale could be retained for use in multiple LMIC and as the basis for a new instrument to measure depression globally. However, the item "loss of sexual interest or pleasure" could be removed as it appears weakly associated with depression across all settings. The item "blaming yourself for things" showed relatively low discrimination in Colombia, Uganda and Rwanda and had overlapping location parameters with other items, suggesting that this item could also be removed from the HSCL for future use. The item related to suicide ideation performed poorly both across and within settings. However, because of the clinical significance of this symptom and its relatively unique location on the latent trait, it is suggested to retain the item, but not necessarily included it when generating summary scores. Because of the heterogeneity in item performance by setting, some items could be removed when the scale is used in certain areas due to particularly poor performance of the item in these areas. For example, the item "thinking everything is an effort" could be considered for removal from the scale when used in Colombia because of its low discrimination in that setting.

Limitations

All of the data used in this analysis came from trauma affected, non-representative populations, limiting the generalizability of these findings to different contexts. While the data included in this analysis were based on validated scales and collected using similar methodology, it is possible that differences in symptom recall timeframes (i.e. 1-4 weeks), idiosyncratic differences of interviewers, differences in recruitment strategies, or other differences in study procedures, may have lead to variability in responses and item parameters. In the present study, we did not explore other variables that could be responsible for the observed DIF, such as gender,

age, education level, or other unmeasured variables. Future studies should explore item performance and DIF related to other potential sources of response bias.

Conclusions

The Hopkins Symptom Checklist 15-item depression scale performed well across diverse settings, with most items showing a strong relationship to the underlying trait of depression. Overall, items were most informative for people with lower than average levels of depression. Some items performed poorly across settings including "loss of sexual interest and pleasure," "thoughts of killing oneself/suicide," and "blaming yourself for things." Almost all of the items on the HSCL showed DIF, however the impact of this DIF was salient in only half of the settings. Future use of the HSCL could investigate potential DIF and adjust for it before comparing scores on the measure across settings. This was the first study to examine the performance of depression related measurement items across multiple settings in LMIC with trauma-affected populations. The methods used in this investigation illustrate the richness of information provided by IRT for scale development and/or refinement. The findings from the study will partially inform the development of a new instrument to measure depression globally.

References

Applied Mental Health Research Group (AMHR). (2013). *Design, implementation, monitoring and evaluation of cross-cultural trauma related mental health and psychosocial assistance programs: A user's manual for researchers and program implementers*. Retrieved from: http://www.jhsph.edu/research/centers-and-institutes/center-for-refugee-and-disaster-response/response_service/AMHR/dime

Baker, F. B. (2001). *The basics of item response theory*. Eric Clearinghouse on Assessment and Education.

Bass, J., Poudyal, B., Tol, W., Murray, L., Nadison, M., & Bolton, P. (2012). A controlled trial of problem-solving counseling for war-affected adults in aceh, indonesia. *Social Psychiatry and Psychiatric Epidemiology, 47*(2), 279-291.

Betancourt, T. S., Yang, F., Bolton, P., & Normand, S. (2014). Developing an african youth psychosocial assessment: An application of item response theory. *International Journal of Methods in Psychiatric Research, 23*(2), 142-160.

Bolton, P., Lee, C., Haroz, E. E., Murray, L., Dorsey, S., Robinson, C., . . . Bass, J. (2014a). A transdiagnostic community-based mental health treatment for comorbid disorders: Development and outcomes of a randomized controlled trial among Burmese refugees in Thailand. *PLoS Medicine, 11*(11), e1001757.

Bolton, P., Wilk, C. M., & Ndogoni, L. (2004). Assessment of depression prevalence in rural uganda using symptom and function criteria. *Social Psychiatry and Psychiatric Epidemiology, 39*(6), 442-447.

Bolton, P., Bass, J. K., Zangana, G., Kamal, T., Murray, S., Kaysen, D., . . . Rosenblum, M. (2014b). A randomized controlled trial of mental health interventions for survivors of systematic violence in Kurdistan, Northern Iraq. *BMC Psychiatry, 14*(1), 360-375.

Bolton, P., Neugebauer, R., & Ndogoni, L. (2002). Prevalence of depression in rural rwanda based on symptom and functional criteria. *The Journal of Nervous and Mental Disease, 190*(9), 631-637.

Cai, L., Du Toit, S., & Thissen, D. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [computer software]*. Chicago, IL: Scientific Software International.

Canel-Çınarbaş, D., Cui, Y., & Lauridsen, E. (2011). Cross-Cultural Validation of the Beck Depression Inventory–II Across US and Turkish Samples.*Measurement and Evaluation in Counseling and Development*, *44*(2), 77-91.

Crane, P. K., Belle, G. v., & Larson, E. B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine, 23*(2), 241-256.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1974). The hopkins symptom checklist (HSCL): A self-report symptom inventory. *Behavioral Science, 19*(1), 1-15.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Ertl, V., Pfeiffer, A., Saile, R., Schauer, E., Elbert, T., & Neuner, F. (2011). Validation of a mental health assessment in an african conflict population. *Psychological Assessment, 22*(2), 318-324.

Fabre, L. F., & Smith, L. C. (2012). The effect of major depression on sexual function in women. *The Journal of Sexual Medicine, 9*(1), 231-239.

Hambleton, R. K., Waminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (1st ed.). California: Sage Publications Inc.

Haroz, E. E., Bass, J. K., Lee, C., Murray, L. K., Robinson, C., & Bolton, P. (2014). Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the Thailand Burma border. *BMC Psychology, 2*(1), 31-40.

Hesbacher, P. T. (1980). Psychiatric illness in family practice. *Journal of Clinical Psychiatry; Journal of Clinical Psychiatry, 41,* 6-10.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Kennedy, S. H., & Rizvi, S. (2009). Sexual dysfunction, depression, and the impact of antidepressants. *Journal of Clinical Psychopharmacology, 29*(2), 157-164.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine*, *32*(6), 959-976.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual review of public health*, *34*, 119-138.

Kitamura, T., Hirano, H., Chen, Z., & Hirata, M. (2004). Factor structure of the zung self-rating depression scale in first-year university students in Japan. *Psychiatry Research, 128*(3), 281-287.

Kojima, M., Furukawa, T. A., Takahashi, H., Kawai, M., Nagaya, T., & Tokudome, S. (2002). Cross-cultural validation of the beck depression inventory-II in japan. *Psychiatry Research, 110*(3), 291-299.

Lizardi, D., & Gearing, R. E. (2010). Religion and suicide: Buddhism, Native American and African religions, atheism, and agnosticism. *Journal of Religion and Health, 49*(3), 377-384.

Mollica, R. F., McDonald, L. S., Massagli, M. P., & Silove, D. M. (2004). *Measuring trauma, measuring torture.* Cambridge, MA: Harvard University.

Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., & Ustun, B. (2007). Depression, chronic diseases, and decrements in health: Results from the world health surveys. *The Lancet, 370*(9590), 851-858.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (Seventh Edition ed.). Los Angeles, CA: Muthén & Muthén.

Nock, M. K., Borges, G., Bromet, E. J., Alonso, J., Angermeyer, M., Beautrais, A., . . . Williams, D. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry, 192*(2), 98-105.

Nuevo, R., Dunn, G., Dowrick, C., Vázquez-Barquero, J. L., Casey, P., Dalgard, O. S., ... & Ayuso-Mateos, J. L. (2009). Cross-cultural equivalence of the Beck Depression Inventory: A five-country analysis from the ODIN study. *Journal of Affective Disorders*, *114*(1), 156-162.

Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the patient-

reported outcomes measurement information system (PROMIS(R)): Depression, anxiety, and anger. *Assessment, 18*(3), 263-283.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R. K. Hambleton. *Handbook of Modern Item Response* Theory (85-100). New York: Springer.

Silove, D., Manicavasagar, V., Mollica, R., Thai, M., Khiek, D., Lavelle, J., & Tor, S. (2007). Screening for depression and PTSD in a cambodian population unaffected by war. *The Journal of Nervous and Mental Disease, 195*(2), 152-157.

StataCorp. (2011). *Stata statistical software* (Release 12 ed.). College Station, TX: StataCorp LP.

Weiss, W., & Bolton, P. (2010). *Assessment of torture survivors in southern Iraq: Development and testing of a locally-adapted assessment instrument.* United States Agency for International Development.

Weissman, M. M., Bland, R. C., Canino, G. J., Faravelli, C., Greenwald, S., Hwu, H., . . . Lellouch, J. (1996). Cross-national epidemiology of major depression and bipolar disorder. *Jama, 276*(4), 293-299.

Wessells, M. G. (2009). Do no harm: Toward contextually appropriate psychosocial support in international emergencies. *The American Psychologist, 64*(8), 842-854.

Whiteford, H., Ferrari, A., Degenhardt, L., Feigin, V., & Vos, T. (2015). The global burden of mental, neurological and substance use disorders: An analysis from the global burden of disease study 2010. *PLoS ONE, 10*(2), e0116820.

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*(1), 42-57.

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339-361.

Development, reliability and validity of the International Depression Symptom Scale (IDSS): A measurement instrument for global presentations of depression

Abstract

Accurate screening of people in need of mental health services in low- and middle-income countries (LMIC) often is done by untrained, non-mental health professionals. Administration of self-report measurement instruments is a common method for screening of common mental health disorders by non-mental health professionals in LMIC. However, these instruments are often based on clinical presentation of western populations and can require intensive resources to adapt and test them in a local context. The aim of this study was to assess the reliability and validity of a self-report scale, the International Depression Symptom Scale-Global version (IDSS-G), that was developed based on empirical evidence of the signs and symptoms of depression that present globally. The IDSS-G showed high internal consistency reliability ($\alpha = 0.92$), test-retest reliability ($r = 0.87$) and inter-rater reliability ($ICC = 0.90$). Construct validity was supported as the IDSS-G was strongly correlated with scores on the Patient Health Questionnaire (PHQ-9), as well as with impaired function and suicidal ideation. Criterion validity was established, and strongest for use of the IDSS-G to identify people with a depressive disorder (Major Depression/Dysthymia). Incremental validity was also supported, as the IDSS-G predicted functional impairment above and beyond what was predicted by the PHQ-9. Overall the results suggest that the IDSS-G is a useful self-report screening measure for depression and may be preferable to the use of a translated scale from western populations.

Introduction

Depression is one of the largest contributors to disease burden worldwide (Ferrari et al., 2013; Murray et al., 2013). In Low and Middle Income Countries (LMIC) the estimated point prevalence of depression ranges from 4.0% in Southeast Asia and South America to 8.6% in South Asia (Ferrari et al., 2012). Despite this significant burden there are few, if any, treatment options in many parts of the world. Only about 2% of people with mental disorders worldwide receive treatment (Wang et al., 2007). There are on average 8.6 psychiatrists per 100,000 people in High-Income Countries (HIC), but this number drops down to 0.05 psychiatrists per 100,000 people in Low-Income Countries (LIC) (World Health Organization, 2011). It has been estimated that it would take the addition of 239,000 full time mental health professionals to the workforce in order to treat all who are in need (Bruckner et al., 2011). This large burden, coupled with limited treatment resources is what is known as the "treatment gap" in global mental health (Lund et al., 2012; Patel, Simon, Chowdhary, Kaaya, & Araya, 2009).

Several strategies have been proposed to help reduce this gap, the most popular one being task-sharing/task-shifting methods. This method requires shifting of tasks related to treating mental health care away from trained psychiatrists/psychologist, to non-mental health specialists such as primary health care workers or community lay workers (Patel et al., 2009; Patel, 2007). Treatment delivered by non-specialist health workers has been shown to reduce symptoms of depression (Bass et al., 2013; Bolton et al., 2003; Bolton et al., 2014; Rahman, Malik, Sikander, Roberts, & Creed, 2008) and increase the number of adults who recover from depressive disorders (van Ginneken et al., 2013).

A necessary first step in treating depression is identifying those individuals with high enough symptoms to need, and likely benefit from treatment. Non-specialist workers do not have extensive training in recognizing the signs and symptoms of mental disorder. As such, they most often rely on the administration of self-report measurement instruments to identify people with depressive illness. Most measurement instruments used in LMIC by non-specialist workers were

originally developed and validated in western-clinical samples (Mulrow et al., 1995) (See chapter 2 for more detail). Many of these have now been adapted and tested in a variety of low-income and non-western settings as well (Adewuya, Ola, & Afolabi, 2006; Ghimire, Chardoul, Kessler, Axinn, & Adhikari, 2013; Haroz et al., 2014a; Patel et al., 2008).

Establishing validity and reliability of an instrument in one setting does not necessarily mean it will be reliable and valid in another. Moreover, the time and resources needed to adapt and test screening instruments for each setting can be cumbersome (Hollifield, 2002). Simple translation and back translation can be rather easy to do, but it does not ensure reliability and validity across settings (Kohrt et al., 2011). Effective methods for establishing the reliability and validity of measurement instruments in a variety of settings, often involve initial qualitative research to identify relevant local problems and terminology, followed by an instrument testing and validation process with relatively large samples (Bass, Ryder, Lammers, Mukaba, & Bolton, 2008; Betancourt et al., 2009; Bolton, 2001; Haroz et al., 2014). However, for many Non-Governmental Organizations (NGOs), Community Based Organizations (CBOs), or primary health centers, who are the main employers of non-specialist mental health care workers, the time and resources needed to properly adapt and test measurement instruments may not be available (e.g. study sample, interviewers/evaluators, study coordinators, data analysts, etc.).

Instead of adapting western measures, some researchers have developed locally relevant screening instruments that are specific to a population. These are designed based on initial qualitative research and consist of signs and symptoms of depression that arise in the population for which it is developed. Examples of these types of instruments include the Shona Symptom Questionnaire (Patel, Simunyu, Gwanzura, Lewis, & Mann, 1997), the Afghan Symptom Checklist (Miller et al., 2006), and the Phan Depression Scale (Phan, Steel, & Silove, 2004). While these instruments benefit from local acceptability, they cannot necessarily be used to make comparisons across contexts. And, as with adaptation of western-based scales, the creation of locally relevant instruments can be resource intensive.

Regardless of the origin of the instrument, instruments to assess for depression can be beneficial for scaling up psychological services and helping non-specialist mental health care workers make critical decisions regarding treatment. Use of valid screening tools to proactively find people in need of treatment can enhance demand for psychological interventions (Patel, Chowdhary, Rahman & Verdeli, 2011). Once cases are identified, decisions regarding appropriate plans for and potential modification of treatment, are made based on this initial screening evaluation. A recent study, in the context of trauma affected populations, has shown that an intervention that allows for modification of treatment based on information obtained from self-scales, is effective for reducing depression symptoms (Bolton et al., 2014; Murray et al., 2013).

Development of the International Depression Symptom Scale

To address the shortage of measures that have been developed for use with non-western populations with the potential to be used across a wide-variety of populations for screening and clinical purposes, research was undertaken to first determine if there was a need for a new instrument, and if needed, develop a the instrument. Two approaches were used in this investigation. The first step involved a systematic review of qualitative research to identify common psychological symptoms related to depression across geographic regions, gender and context. The second step was a quantitative analysis using Item Response Theory (IRT) of data on the 15-symptoms of the HSCL-25 depression scale (Hesbacher, 1980; Mollica, McDonald, Massagli, & Silove, 2004; Winokur, Winokur, Rickels, & Cox, 1984) from eight distinct cultural settings. The IRT analysis was done to evaluate the performance of the different symptoms across populations.

Results from the literature review suggested that existing scales do not include all of the most common symptoms of depression that occur globally (see Chapter 3). Results from the quantitative analysis indicated that several items on the HSCL perform poorly across and within

settings (see Chapter 4). Taken together, the results suggest that a new instrument should be created that better reflects global presentations of depression.

To do so, symptoms that were common across multiple regions identified from the literature search were combined with the best performing symptoms from the quantitative analysis to create a draft instrument. An expert panel of 7 researchers and practitioners from the field of global mental health, anthropology, psychiatric epidemiology, and psychiatry reviewed this draft instrument and additional revisions were made based on their feedback. These revisions included the addition of two items from the DSM diagnostic criteria for Major Depressive Disorder that were not already included in the first draft of the instrument (*psychomotor agitation and slowing* and *problems with concentration*).

The resulting draft instrument is titled the 'International Depression Symptom Scale (IDSS) (Appendix E). The IDSS is intended to be used as a modular instrument, whereby 29 items make up the overall global measure (IDSS global version; IDSS-G) and additional items are added in when it is used in different locations (IDDS local version; IDSS-L). Which additional items are included can be based on frequencies of symptoms arising from the literature review (i.e. was a particular symptom very common in some populations but not others) as well as any previous qualitative work that has identified particularly relevant symptoms in one setting that are not already included on the global measure.

The IDSS is designed to either be used as a self-report instrument to be filled out by the participants themselves or to be administered by a local interviewer.Scoring of the IDSS is done in two ways. For the global measure (IDSS-G) an average of the first 27 items is calculated. The item that relates to functional impairment ("difficulty doing your usual activities at home or work") and the item that relates to suicidal ideation ("thoughts of wanting to kill yourself") are not intended to be included in summary scores as these are for clinical information purposes only. A locally specific score (IDSS-L) is also calculated by generating average scores across all 27 items on the IDSS-G, as well as any additional locally specific items included. The 29 symptoms

on the IDSS-G, the additional 1 symptom added as part of the IDSS-L based on previous research indicating its local importance, and the supporting development process that informed each symptom's inclusion is provided in Table 1.

Table 6.1

Source of the supporting evidence for each item on the IDSS

Item	Qualitative review	IRT analysis	DSM-5
D01 sad	◆	◆	◆
D02 no interest	◆	◆	◆
D03 crying	◆	◆	
D04 hopeless	◆	◆	
D05 lonely	◆	◆	
D06 social withdrawal	◆		
D07 tired/fatigue	◆	◆	◆
D08 weigh too little	◆	◆	◆
D09 weigh too much	◆	◆	◆
D10 increased appetite	◆	◆	◆
D11 sleep problems	◆	◆	◆
D12 feeling trapped		◆	
D13 worry	◆	◆	
D14 worthless	◆	◆	
D15 headaches	◆		
D16 stomachaches	◆		

D17 general aches and pains	◆		
D18 anger	◆		
D19 thinking too much	◆		
D20 confused	◆		◆
D21 heart weakness	◆		
D22 palpitations	◆		
D23 heavy heart	◆		
D24 heart pressure	◆		
D25 heart pain	◆		
D26 psychomotor			◆
D27 concentration			◆
D28 disappointed[a]	◆		
D29 imp function			◆
D30 suicide			◆

[a] Only included in the IDSS-L. This symptom was included based on its frequency in Southeast Asian populations arising form the literature review as well as previous qualitative work among Burmese refugees living in Thailand

Current Study

The aim of the current study was to test the reliability, validity and clinical utility of the IDSS-G in a community sample of adults in Yangon, Myanmar. In addition, to assess if the IDSS-G is an improvement on an un-adapted, but translated, measurement instrument, the ability of the IDSS-G to identify people with mental health problems above and beyond a non-adapted measure for depression (PHQ-9), was evaluated (i.e. *incremental* validity). Results in the present study are limited to the IDSS-G, but results examining the performance of the IDSS-L (IDSS-G

plus one additional item, “disappointment”) can be found in Appendix F. This study is the first step in the broader goal of having a reliable and valid instrument to measure depression that better reflects global commonalities in symptoms of depression across and within, different cultural contexts.

Methods

Measures

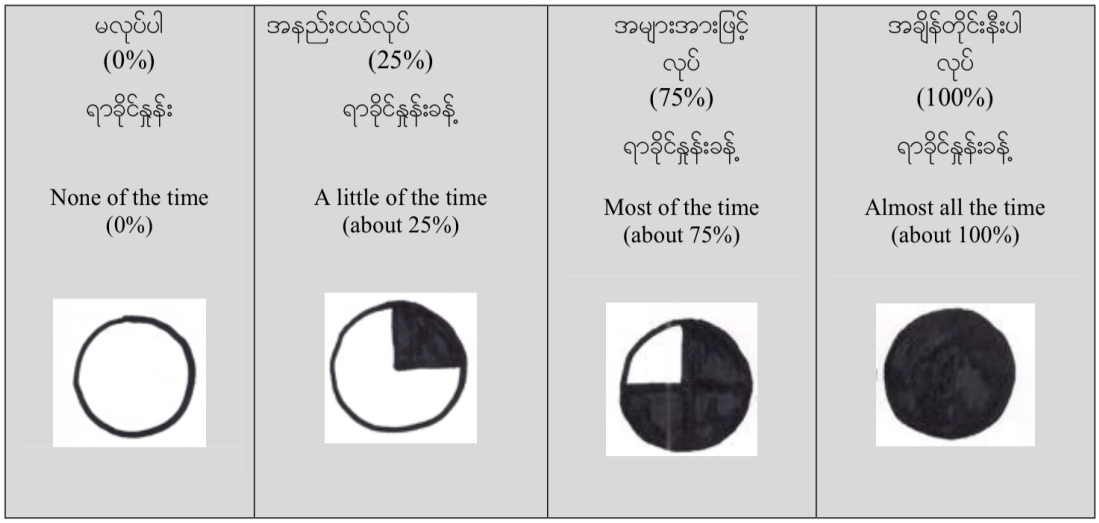
Assessment battery

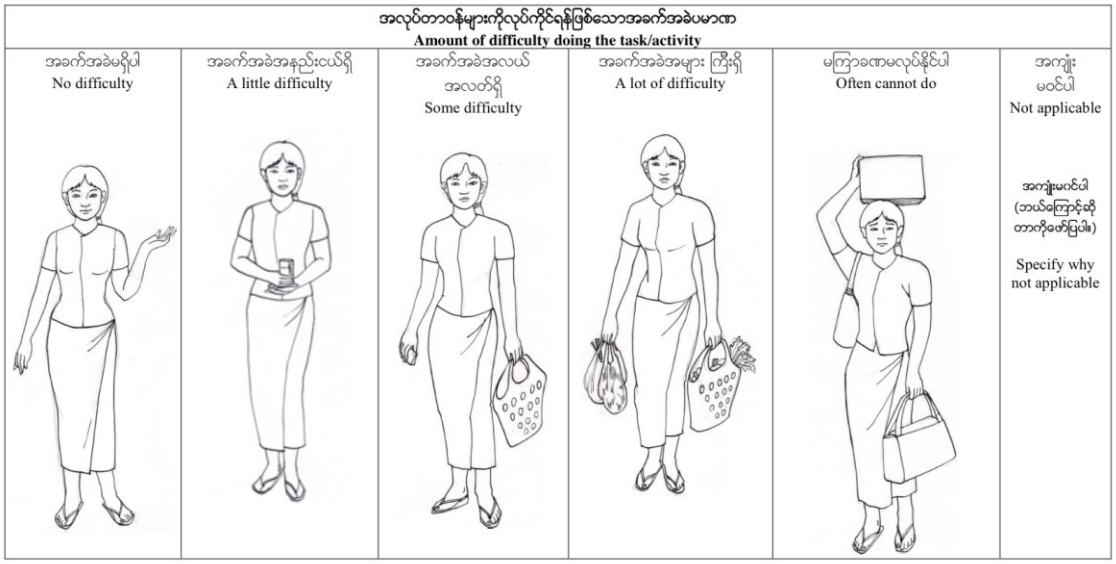
Eight local interviewers, collected data via mobile data collection device on a number of measures as part of the assessment battery. Basic demographics, including age and sex, were also collected.

The International Depression Symptom Scale Global version (IDSS-G) is a 29-item self-report measure. It is designed for either self-report of symptoms by individuals or for administration by a local interviewer. Participants were asked to indicate how often in the last two weeks they had experienced each symptom in the measure. Responses options ranged from 0 “none of the time” to 3 “almost all the time.” To assist participants in interpreting the response options, a visual cue card representing each option as a percent of a circle, was used (Figure 1).

Figure 6.1

Example visual cue cards for response options on the IDSS-G and impaired functioning measure





Patient Health Questionnaire-9 (PHQ-9; Kroenke & Spitzer, 2002) is a 9-item self-report measure that uses likert-type response options. Participants are asked how often in the past two weeks had the symptom bothered him/her and response options ranged from 0 “not at all” to 3 “nearly every day.” The PHQ-9 is a commonly used measure of depression and has been found

to be valid in other low-resource settings such as Haiti(Marc et al., 2014), Peru (Zhong, Gelaye, Fann, Sanchez, & Williams, 2014), and Thailand (Lotrakul, Sumrithe, & Saipanish, 2008). The PHQ-9 has not been previously validated for use in Yangon, Myanmar.

Local measure of functional impairment. The functional impairment measure was developed based on a previous qualitative study involving Burmese refugees displaced in Thailand and validated for that population (Haroz et al., 2014). The measure includes tasks and activities that were identified by the Burmese refugees as important for men and women to do in order to care for themselves, their families and their communities. There are separate scales for men (16 items) and women (23 items). For each item participants were asked about how much more difficulty he/she has had in the last four weeks in doing each task and activity compared to other men/women of similar age. Response options included 0 "no difficulty" 1 "a little difficulty" 2 "some difficulty" 3 "a lot of difficulty" and 4 "often cannot do." To help respondents with interpreting the response options, a visual cue card depicting a man or a woman struggling to carry varying amounts of bricks/groceries was used (Figure 1). While not previously validated with the current study sample, at the time of the present study, the instrument was being used in another study in Yangon with no reported problems of its relevance.

Structured clinical interview

Structured Diagnostic Interview for DSM-IV (SCID; (First, Spitzer, Gibbon, & Williams, 2012) is a semi-structured interview designed for use by trained mental health professionals to facilitate making the major Diagnostic and Statistical Manual Axis I disorder diagnoses (DSM-IV; (American Psychiatric Association, 2000). The SCID includes specific structured questions to help elicit diagnostic information, but ultimately it is up to the mental health professional administering the interview to decide whether each criterion for each disorder is met. For the current study, three clinical diagnoses from the SCID were evaluated: Major Depressive Episode/Disorder (MDD), Dysthymia, and Generalized Anxiety Disorder (GAD). Although the

study was mostly concerned with MDD and Dysthymia, GAD was included as a possible diagnosis due to its high co-morbidity with depression (Almeida et al., 2012) and overlap in diagnostic criteria (American Psychiatric Association, 2013).

During training, the psychiatrists were given a handout with the DSM-IV criteria for each of the 3 study disorders (MDD, Dysthymia, and/or GAD) and were instructed to use the SCID as a way to ask questions and help inform their clinical judgment about whether the person met the DSM criteria for any of the relevant disorders or none of them. Following completion of the interview with the IDSS-G every participant was evaluated by a local psychiatrist using an electronic version of the SCID. The first 40 study participants were interviewed by psychiatrists in pairs in order to establish inter-rater reliability, while the remainder of participants were interviewed by psychiatrists working individually.

Translation

The IDSS-G, PHQ-9 and SCID, were translated by the local study coordinator (Burmese woman) with the IDSS-G back translated by a local psychiatrist. No translation was needed for the measure of functional impairment, as it was already in Burmese related to its use in a previous study. Review of the translations for the full assessment battery (PHQ-9, IDSS-G, and functional impairment measure) and SCID took place as part of the training of interviewers and psychiatrists. During training the interviewers and psychiatrists commented on the translation and the meaning of each item. No major problems with translations were identified. However, several problems with the study measures did arise. These problems related to the order in which measures were asked (i.e. functional impairment before IDSS-G), numbering of items, and problems with the mobile data collection process (i.e. psychiatrists wanted SCID answer choices to be in English rather than Burmese, problems with logging into data collection software accounts). All of these problems were corrected prior to data collection.

Interviewers & Psychiatrists

Eight local interviewers administered the full assessment battery, including the IDSS-G, using mobile devices, and facilitated the pile sort activity and cognitive interviewing (described below). Interviewers were people from the local community who were literate in Burmese and had previous experience doing data collection. Interviewers were trained in study procedures, the signs of active psychosis and major developmental delay (i.e. the exclusion criteria), research ethics, and a safety protocol, during a three-day training in Yangon prior to the start of data collection.

Four local psychiatrists conducted the clinical interviews using the SCID. All psychiatrists had medical degrees obtained from medical schools in Myanmar. Three had been practicing psychiatry for more than five years while the fourth was currently finishing residency in psychiatry. All psychiatrists attended a three-day training in Yangon on the SCID, study logistics and procedures, and a safety protocol for adverse events prior to the start of data collection.

Participants

Participants were recruited in partnership with two local medical clinics in Yangon, Myanmar. These clinics provide both general medical and psychiatric care to the community. The average daily patient load of each clinic ranged from 30-40 patients who present with medical complaints, alcohol abuse problems and psychiatric problems. To be included in the study, participants had to be a patient at either of the clinics, literate in Burmese, and over the age of 18. Exclusion criteria consisted of active psychosis or the presence of a major developmental delay as determined by the interviewers who were trained on signs of these conditions as part of training.

Clinic attendees were provided with an information sheet by the clinic staff describing the study and then were asked whether they agreed to be contacted by the study team. If they agreed to be contacted, an interviewer called the participant to set up a meeting to describe the study and administer consent.

All participants provided informed verbal consent for their participation in the study and all study procedures and forms were approved by the Johns Hopkins Bloomberg School of Public Health Internal Review Board (IRB #6011) and the Ethics Review Committee of the Department of Medical Research (Lower Myanmar), Ministry of Health, Republic of Myanmar.

Study procedures

A local interview administered the assessment battery by asking each question to the participant and recording responses on mobile devices. Participants were then asked to participate in a pile sort activity using the symptoms from the IDSS-G; a cognitive interview focused on specific items from the IDSS-G; and/or a re-interview 2-5 days after the first interview.

Pile Sort: Fifty participants were asked to participate in the pile sort activity. The pile sort activity was used as a form of qualitative factor analysis to compare the way respondent's group symptoms together. Participants were given cards with each symptom from the IDSS-G (n = 29) and were asked to put together piles of symptoms that they thought went together. After grouping all the cards, participants were asked to title the piles and explain their reasoning for grouping the symptoms in the way they did. The frequency of each titled grouping, as well as the frequency of each symptom within these groupings, was calculated.

Cognitive interviewing: A subset of participants (n = 30 men and n = 30 women) was asked to complete a cognitive interview to assess face validity and the comprehension of select items from the IDSS-G. Cognitive interviews were done by a pair of interviewers: one to ask the questions and one to write down the responses. For each symptom question, participants were asked: 1) *Please describe the meaning of this question in your own words. Please use examples to help describe the meaning*; 2) *Is there any part of this question you don't understand or that does not make sense?*; 3) *Can you tell me what thought you had when deciding your answer choice? I'd like to know anything you thought of between when I asked you the question and when you gave me your answer*; and 4) *Was this question easy or difficult to answer? If difficult, please tell*

me why it was difficult? Several of the items (n = 16) on the IDSS-G had been previously tested in a similar population (see (Haroz et al., 2014a)) and were not asked about during the cognitive interview. This left a total of n = 13 items from the IDSS-G that were part of the cognitive interviews.

Re-interview: To assess test-retest and inter-rater reliability, n = 54 study participants were interviewed a second time approximately 2-5 days after their first interview. The re-interview was done by the same interviewer who administered the initial interview (n = 24) or a different interviewer (n = 30).

Analysis

Average summary scores for the IDSS-G, PHQ-9 and the local measure of functional impairment, were generated by taking the mean of the item responses on each instrument.. For the IDSS-G, the item that relates to functional impairment ("difficulty doing your usual activities at home or work") and the item that relates to suicidal ideation ("thoughts of wanting to kill yourself") were not included in summary scores, but provided clinical information. All analyses were done using STATA-13 (StataCorp, 2011) and Mplus 7.3 (Muthén & Muthén, 1998-2012).

Reliability

Reliability of the IDSS-G was assessed using: 1) internal consistency reliability, 2) test-retest reliability, and 3) inter-rater reliability. Cronbach's alpha (α) and examination of item-test correlations, item-rest correlations and average inter-item covariance, were used to measure internal consistency reliability of the IDSS-G. Pearson's correlation coefficients (r) were calculated for test-retest reliability comparing average scores on the IDSS-G from the first interview to the average scores on re-interview (done by the same interviewer). Correlations of |0.7| or above are considered very strong, correlations of |0.4| to |0.69| are considered strong, |0.3| to |0.39| are moderate, |0.2| to |0.29| are weak, and anything less than |0.2| are considered negligible (Cohen, 1988). Inter-rater reliability was assessed using intra-class correlation (ICC)

by comparing average IDSS-G scores separately from the first interview to the IDSS-G score on re-interview (done by a different interviewer). Intra-class correlations greater than 0.75 are considered excellent; 0.40-0.75 are considered fair to good; and less than 0.40 considered poor (Fleiss, 1986)..

To establish the reliability of psychiatrist diagnosis, inter-rater reliability between pairs of psychiatrists was calculated using a Kappa statistic. A Kappa of less than 0 indicates less than chance agreement; Kappa of 0.01-0.20 is indicative of slight agreement; 0.21-0.40 indicates fair agreement; 0.41-0.60 indicates moderate agreement; 0.61-0.80 indicates substantial agreement and 0.81-0.99 indicates almost perfect agreement (Viera & Garrett, 2005).

Validity

Validity was assessed through examination of face validity, construct validity, item-level convergent validity, criterion validity and incremental validity. A measure is considered to have face validity when it appears to be measuring what it purports to measure. Construct validity is defined as the degree to which a scale measures the theoretical construct that it was designed to measure and is correlated to other related constructs. Item-level convergent validity in this case referred sufficiently high correlations of items to factors (as measured in the EFA) indicating that the items within a certain factor are highly related. Criterion validity is defined as the association of a scale to a criterion variable (in this case psychiatric diagnosis of DSM disorder) (Allen & Yen, 2002). Incremental validity refers to the ability of the IDSS-G to increase predicative ability beyond existing measures of depression (Sackett & Lievens, 2008).

Face validity was examined during psychiatrist training and as part of the cognitive interviews, by asking participants to describe the meaning of the items, the relevance of these items to depression and in the case of participants, their thought process when answering the items. Construct validity was assessed at the scale and item level. At the scale level Pearson's correlation coefficients (*r*) and Spearman's correlation coefficients (*Rho*; used for correlations

between continuous variables and ordered categorical variables) between average scores on the IDSS-G and age, gender, average scores on the impaired functioning measure, average scores on the PHQ-9, as well as the single items related to functional impairment and suicidal ideation included on the IDSS-G, were calculated. Based on evidence in the literature, it was hypothesized that higher scores on the IDSS-G would be associated with increasing age (Bromet et al., 2011; Jorm, 2000; Kessler et al., 2003); female gender (Bromet et al., 2011; Nolen-Hoeksema, Larson, & Grayson, 1999); worse impaired functioning (Kessler & Bromet, 2013; Ormel et al., 2008), and any suicidal ideation (Nock et al., 2008). As both the IDSS-G and the PHQ-9 measure symptoms of depression, it was expected that scores on these measures would be highly correlated. Evidence for these hypothesized associations in the present study would support the construct validity of the IDSS-G.

At the item level, to investigate construct validity and whether any of the symptoms on the IDSS-G were particularly associated with impaired functioning, 27 single linear regression models were run. Each model included age, gender and one of the symptoms on the IDSS-G. Because of the risk of a Type I error due to multiple comparisons, significance was set using a Benjamini-Hochberg adjusted *p*-value based on a false discovery rate approach (Benjamini & Hochberg, 1995). After correction for multiple comparisons, it was expected that most individual symptoms would be associated with impaired functioning as all of the items on the IDSS-G were thought to be measuring the same underlying latent trait of depression, and thus would partially account for some of the variance in impaired functioning.

In addition, principal components analysis (PCA) and exploratory factor analysis (EFA) with geomin factor rotation were used to assess item-level convergent validity. Convergent validity would be supported if the items within each factor were highly correlated. The PCA was used to help guide the EFA. Based on the results from the PCA, several factor models were compared in the EFA by looking at factor loadings, item uniqueness, and the overall fit of the models (using Fit indices: RMSEA, and CFI, TLI). RMSEA values lower than 0.05 and TLI/CFI

values above 0.90 are indicative of good model fit (Hu & Bentler, 1998). The pile sort activity was used as a form of qualitative factor analysis to compare the way respondent's group symptoms together to the results of the EFA

To explore criterion validity, diagnoses by a psychiatrist as to whether the person was currently suffering from MDD, Dysthymia, and/or GAD (collectively referred to as the SCID disorders) were used. As a first analysis, average scores on the IDSS-G of participants who were diagnosed with any of the SCID disorders were compared to average scores on the IDSS-G for those participants who were identified as having none of the SCID disorders. This was followed by a disorder-specific analysis where average scores on the IDSS-G among those who were diagnosed with Depression or Dysthymia were compared to average scores for those classified as having none of the SCID disorders; and average scores for those classified as GAD were compared to average scores for those classified as having none of the SCID disorders.

Because the IDSS-G was created to reflect global presentations of depression, it was hypothesized, that the IDSS-G would be better at detecting MDD and/or Dysthymia than at detecting GAD. However, because of the strong association of depression disorders and GAD, the IDSS-G might also detect some GAD cases. Thus, criterion validity would be supported if average scores on the IDSS-G were statistically significantly higher among participants with any diagnosis (MDD, dysthymia and/or GAD) and/or a depression disorder (Depression/Dysthymia) compared with participants identified as having none of the SCID disorders. Determination of whether the difference of means between diagnostic categories was statistically significantly was done using paired sample t-tests.

To see if the IDSS-G could distinguish between anxiety and depression related disorders, average scores between participants diagnosed with GAD were compared to participants diagnosed with MDD and/or Dysthymia. Discrimination would be established if the average scores on the IDSS-G are higher for those diagnosed with MDD or Dysthymia than for those

diagnosed with GAD. Determination of whether the difference of means between diagnostic categories was statistically significant was again done using paired sample t-tests.

Incremental validity

Incremental validity was assessed through the use of linear regressions to predict functional impairment as a function of age, suicidal ideation, PHQ-9 scores and average scores on the IDSS-G. An initial regression was performed examining the impact of age on functional impairment (model 1), followed by a regression examining the impact of age and suicidal ideation (model 2); followed by a model including age, suicidal ideation and scores on the PHQ-9 (model 3). In the final regression analysis (model 4), scores on the IDSS-G were included with the other 3 covariates (age, suicidal ideation and PHQ-9 scores). Incremental validity would be supported if scores on the IDSS-G significantly predicted functional impairment ($p < 0.05$), above and beyond the impact of age, suicidal ideation and scores on the PHQ-9.

Clinical Utility

Finally, Receiver Operating Curves (ROC) were used to compare the area under the curve (AUC), a measure of diagnostic utility, for the IDSS-G and PHQ-9 across two of the diagnostic comparisons (no disorder vs. any disorder & MDD/Dysthymia). ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity), for various cutoff values of a continuous measure compared to a dichotomous criterion. An AUC of 0.5 (50% sensitivity and 50% specificity) indicates that the test is of no diagnostic utility, while an AUC of 1.0 (100% sensitivity and 100% specificity) indicates the test under investigation perfectly predicts the criterion. A general guide to interpreting AUC values is that 0.50-0.70 indicates low accuracy; 0.70-0.90 indicates moderate accuracy, and above 0.90 indicates high accuracy (Fischer, Bachmann, & Jaeschke, 2003). In addition, sensitivity, specificity, positive predictive value, negative predictive value, and percent correctly classified, were explored for various cut-

off points on the IDSS-G and PHQ-9. For both measures, optimal cut-off points were generated based on maximizing the sensitivity and specificity for each diagnostic category (Liu, 2012).

Results

Descriptive statistics

Overall $N = 151$ people were interviewed using the IDSS-G and associated measures; $n = 2$ of these individuals refused to participate in the SCID evaluation and $n = 2$ had data that was mistakenly erased during uploading; leaving a final analytic sample of $n = 147$. Two-thirds of the participants were women ($n = 95$; 63.8%) and ages ranged from 18-81 with a mean age of 47.5 (Table 2).

Table 6.2

Demographic information for instrument testing sample (N = 147)

Gender, *n* (%)	
Men	52 (35.4)
Women	95 (64.6)
Age, *M* (SD), *Range*	47.6 (13.6), *18-81*

Table 3 displays the summary statistics for the assessment measures used in the study. The average score on the IDSS-G ranged from 0-2.44 with a mean of 0.72 ($SD = 0.49$). The distribution of scores shows that all of the measures were positively skewed, indicating that most participants reported few symptoms and good functioning (Figure 2).

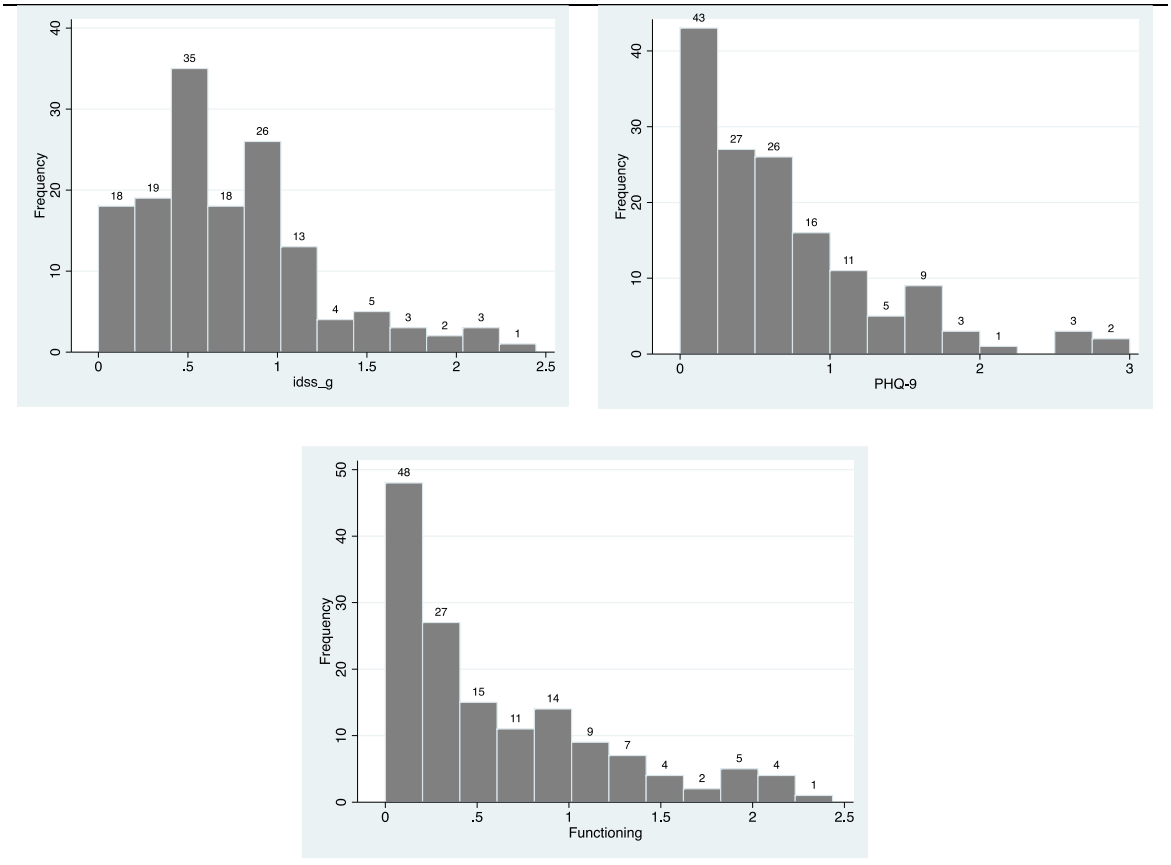
Table 6.3

Mean scores and frequencies for each measurement instrument used in assessment battery

Measure	*N*	*M*	*Range*	*SD*	*Skew*
IDSS-G	147	0.72	0-2.44	0.49	1.07
PHQ-9	146	0.67	0-3	0.63	1.46
Functioning	147	0.61	0-2.43	0.60	1.08

Figure 6.2

Histograms of summary scores on the IDSS-G, PHQ-9, and impaired functioning measure



Based on psychiatrist diagnosis with the SCID, *n* = 31 people met diagnostic criteria for Major Depressive Disorder (MDD), *n* = 39 people for Dysthymia, and *n* = *22* for Generalized Anxiety Disorder (GAD) (Table 4). Thirty participants only met criteria for a single diagnosis and *n* = 24 met criteria for a comorbidity. Comorbidities included: *n* = 18 with MDD and Dysthymia; *n* = 3 with MDD and GAD; and *n* = *2* with Dysthymia and GAD. One participant was diagnosed with all three disorders (Table 4). A little less than half of the total sample (*n* = 63; 42.9%) were classified as having none of the SCID disorders.

Table 6.4

Frequency of SCID based DSM diagnoses (N = 147)[a]

	N (%)
Any disorder	71 (48.3)
Depression	31 (21.1)
Dysthymia	39 (26.5)
GAD	22 (15.0)
None of these disorders	63 (42.9)
Co-morbidity (2 or more)	24 (16.3)

[a] Some individuals who were part of the analytic sample were diagnosed as having PTSD, but PTSD diagnoses were not included in the criterion validity analysis.

Reliability Results

Internal Consistency Reliability and item analysis

Cronbach's Alpha (α) calculations, item-test correlations, item-rest correlations and average inter-item covariance for the IDSS-G are presented in Table 7. The Cronbach's alpha was

high for the IDSS-G ($\alpha = 0.92$). Analysis of item level correlations supported dropping the item "weighing too much" as the item was negatively correlated with all other items (Table 7).

Table 6.7

Item analysis of the items on the IDSS-G[a]

	# of obs	Sign	Item-test correlation	Item-rest correlation	Average inter-item covariance	Alpha of scale if item is removed
D01 Sad	147	+	0.7357	0.6987	.2120844	0.9117
D02 no interest	144	+	0.5448	0.4950	.2210306	0.9153
D03 crying	146	+	0.6674	0.6327	.2192119	0.9134
D04 hopeless	146	+	0.6275	0.5853	.2186875	0.9139
D05 lonely	147	+	0.5943	0.5480	.2193612	0.9145
D06 social withdrawal	147	+	0.6106	0.5638	.2181047	0.9142
D07 tired/fatigue	147	+	0.6597	0.6165	.2159656	0.9133
D08 weigh too little	143	+	0.5661	0.5267	.2233151	0.9150
D09 weigh too much	142	-	0.0876	0.0526	.2369992	0.9198
D10 increased appetite	147	+	0.5108	0.4608	.2228915	0.9158
D11 sleep problems	147	+	0.5857	0.5302	.2170149	0.9149
D12 trapped	147	+	0.7307	0.6943	.2131216	0.9119
D13 worry	147	+	0.6322	0.5842	.2162514	0.9139
D14 worthless	147	+	0.5629	0.5175	.2215674	0.9150
D15 headaches	147	+	0.5509	0.4944	.219151	0.9156
D16 stomach_aches	147	+	0.1816	0.1284	.2349875	0.9203
D17 other_aches	147	+	0.4755	0.4143	.2225897	0.9171
D18 anger	147	+	0.4745	0.4264	.2251278	0.9163
D19 thinking too much	147	+	0.6229	0.5695	.2148866	0.9142

D20 confused	147	+	0.6774	0.6396	.2168961	0.9129
D21 heart_weakness	147	+	0.5155	0.4615	.2218682	0.9160
D22 palpitations	146	+	0.6334	0.5934	.2190189	0.9137
D23 heavy_heart	146	+	0.6076	0.5686	.2210161	0.9142
D24 heart_pressure	146	+	0.6811	0.6495	.2195777	0.9132
D25 heart_pain	146	+	0.5016	0.4616	.2258162	0.9158
D26 psychomotor	145	+	0.5315	0.4863	.2229376	0.9154
D27 concentration	146	+	0.4920	0.4347	.2220678	0.9164
Whole Scale					.2207999	0.9179

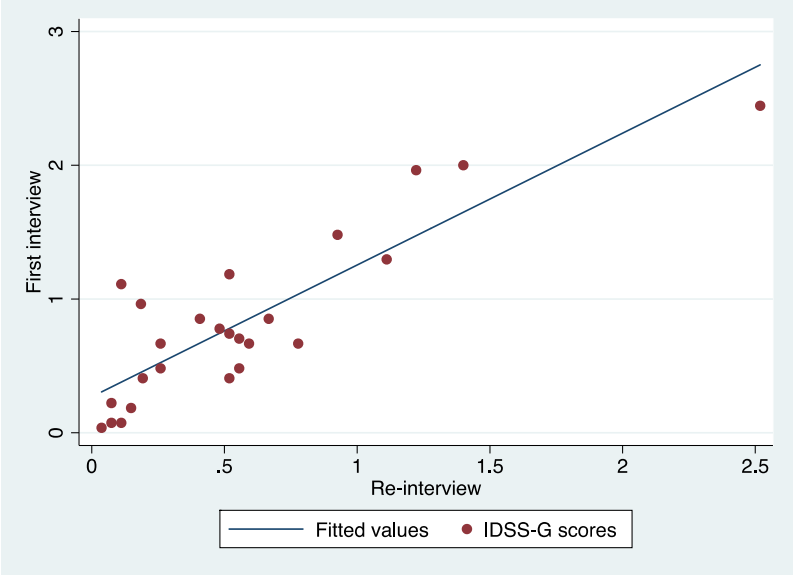
[a] The impaired functioning and suicidal ideation items were not included in the Cronbach's alpha analysis, since these items are not intended to be included in summary scores.

Test-Retest Reliability

Initial interviews and re-interviews with the same interviewer were done with $n = 24$ participants. Re-interviews were performed within 2-11 days of the initial administration of the IDSS-G, and the average time between interviews for test-retest reliability was 3.8 days ($SD = 2.17$). Visual inspection of the graph depicting the relationship between average scores on the IDSS-G at the first interview and re-interviews indicated that a linear relationship fit the data well (Figure 3). The correlation between average scores on the first interview with average scores on the re-interview was $r = 0.87$, indicating a strong positive relationship and good test-retest reliability.

Figure 6.4

Scatter plot of average IDSS-G scores at baseline and re-interview



Inter-Rater Reliability

For the IDSS-G, initial interviews and re-interviews with different interviewers were done with n = 30 individuals. On average re-interviews were done 10.2 days (SD = 5.3), with a range of 2-19 days, after the initial administration of the IDSS-G. The ICC across interviewers for average score on the IDSS-G was ICC = 0.90 with a 95%CI of [0.79, 0.95], indicating high inter-rater reliability.

For the psychiatric diagnoses, the Kappas and percent agreements for each pair of psychiatrists is presented in Table 8. Pair 1 jointly rated n = 16 participants and Pair 2 jointly rated n = 26 participants. For both pairs, the Kappas indicated that for all diagnoses there was substantial to almost perfect agreement, with the exception of the Dysthymia rating in Pair 1 for which only fair agreement was achieved.

Table 6.8

Inter-rater reliability using Kappa statistic by pair of psychiatrist

Criterion	Pair 1 (n = 16)	Pair 2 (n = 26)
	Kappa (% agreement)	Kappa (% agreement)
MDD vs. all other diagnoses	0.86 (93.8)	0.78 (96.2)
Dysthymia vs. all other diagnoses	0.38 (68.8)	0.75 (92.3)
GAD vs. all other diagnoses	0.85 (93.8)	1.00 (100.0)
No diagnosis vs. all other diagnoses	0.64 (93.8)	0.91 (96.2)

Validity

Face validity

Cognitive Interviews

Cognitive interviewing was done to assess the face validity and comprehensibility of n = 13 items on the IDSS-G. Sixty participants (n = 30 female, n = 30 male) completed cognitive interviews. Most questions were easily understood, with the exception of "feeling weakness in your heart" [n = 15 found it difficult to understand]; "feeling as though your heart was heavy" [n = 7]; "pain in your heart" [n = 1]; and "difficulty concentrating" [n = 1].

Appendix G lists the items that were included as part of the cognitive interviews and the top 3 most frequent responses related to the meaning of each of the items. Many of explanations were not explicitly related to mental health. For example, the majority of people talked about "stomach pain" being related to medical problems or eating spicy food. Only one person mentioned that stomach pain could come from stress. The item "other bodily aches and pains"

also overwhelmingly was reported to be related to physical and medical issues, with most respondents describing having this symptom after being sick or having a medical issue [n = 29], working too much [n = 15], or being caused by cold weather [n = 14]. The meanings of the items "feeling weakness in your heart," "heart palpitations" "feeling pressure on your heart" and "pain in your heart" were described as related to medical problems as well.

The item "Thinking too much" was mostly related to thinking about the current political situation in Myanmar, family, or the future. A few participants talked about "thinking too much" as related to stress [n = 2], stating: "thinking a lot is when you're having stress."

Most participants indicated that the question "moving or speaking so slowly or so fast that others have noticed," was easy to understand and answer, but the responses about the meaning of this item had little to do with mental health. The most frequent response for the meaning of this item was "This is a medical problem and the person is probably tired or overly excited" [n = 12] and "When I am talking about something that I like, my talking is fast" [n = 12]. Only two people reported a connection of this symptom to thoughts, feelings or behavior: "My coworkers have told me that I am very slow at work, all my movements are slow and I am slow in doing my tasks. I think this is because I do not have enough motivation to do my work" and "When I am worried or anxious I cannot control my behavior, so I might move or speak slowly without noticing that"

Construct validity

Table 11 displays the polychoric correlation matrix for the scores on the: 1) IDSS-G; 2) age; 3) gender; 3) functional impairment measure; 4) PHQ-9; 5) functional impairment item; and 6) the suicidal ideation item. The IDSS-G is not significantly correlated with age and gender. Construct validity was supported by a very strong correlation between the IDSS-G and the PHQ-9 (r = 0.78) and strong correlations between the IDSS-G and functional impairment (r = 0.56), single functional impairment item (*Rho* = 0.65), and single suicidal ideation item (*Rho* = 0.65).

The correlation between the functional impairment measure and the single functional impairment question was strong as well (*Rho* = 0.48) indicating that while a single indicator may not fully capture all aspects of impaired functioning, it could serves as a brief indicator of this domain.

Table 6.11

Exploration of construct validity: Correlations of IDSS-G and other measured variables

	IDSS-G	Age	Gender	Functioning Measure	PHQ-9	Function Item	Suicide Item
IDSS-G	1.00						
Age	-0.16	1.00					
Gender	0.17	-0.06	1.00				
Functioning Measure	0.56*	-0.17*	-0.11	1.00			
PHQ-9	0.78*	-0.18*	0.06	0.50*	1.00		
Functioning Item	0.65*	-0.16	-0.05	0.48*	0.62*	1.00	
Suicide item	0.65*	-0.40*	0.09	0.50*	0.56*	0.56*	1.00

* $p < 0.05$

For item-level construct validity, all but four items significantly predicted more impairment in functioning after adjusting the *p*-value to account for multiple comparisons. The items that significantly predicted the most impaired functioning were "crying a lot" ($\beta = 0.33$) and "weighing too little" ($\beta = 0.31$). The effects of the other significant items ranged from $\beta = 0.16$ to $\beta = 0.28$. The items "weighing too much," "stomach aches," "other aches and pains" and "pain in the heart" did not significantly predict impaired functioning.

Convergent validity

Factor Analysis

The PCA indicated one predominant factor with an eigenvalue of 10.19. There were five other factors with eigenvalues above 1 (1.99, 1.51, 1.41, 1.21, 1.15 respectively) (Figure 3). Based on this information 1, 2, 3, 4, and 5 factor solutions were explored using EFA, comparing model fit statistics (Table 5) and looking at what made theoretical sense. The three-factor solution was selected as the most appropriate model and model fit statistics indicated good model fit (RMSEA = 0.047; CFI = 0.959; TLI = 0.949).

Figure 6.3

Scree plot with parallel analysis for items on the IDSS-G

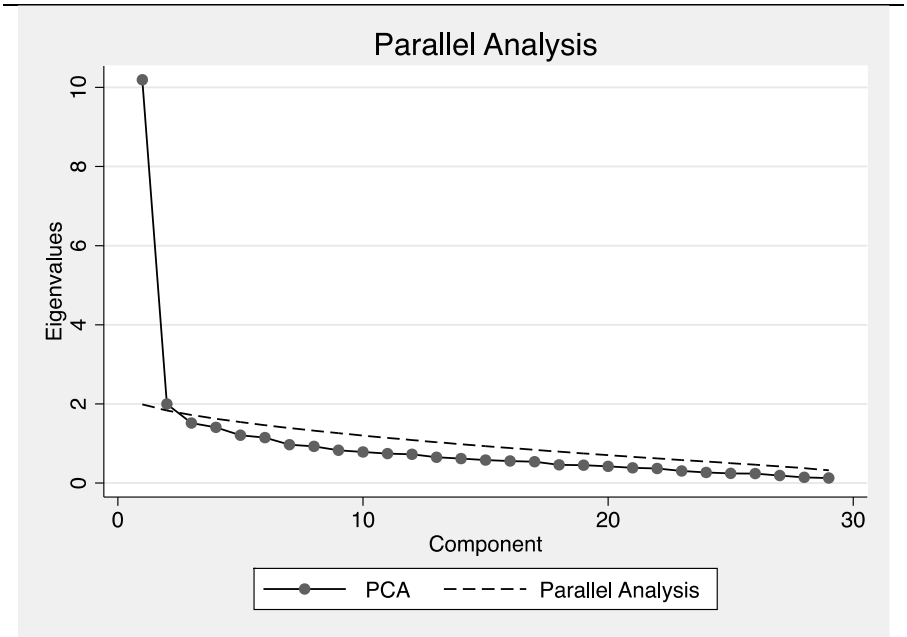


Table 6.5

Model fit indices and their standard errors for various factor solutions for the IDSS-G[a]

Model	X^2	*df*	P value	RMSEA	CFI	TLI
1 factor model	629.37	377	0.000	0.067	0.903	0.896
2 factor model	503.72	349	0.000	0.055	0.941	0.931
3 factor model	428.08	322	0.000	0.047	0.959	0.949
4 factor model	371.28	296	0.002	0.042	0.971	0.960
5 factor model	331.04	271	0.007	0.039	0.977	0.966

[a]RMSEA values lower than 0.05 and TLI/CFI values above 0.90 are indicative of good model fit

The 3-factor model was run with geomin rotation to generate factor loadings for each item (Table 6). The majority of the items loaded on the first factor, and include symptoms related to depressed mood, social isolation, and cognitive impairment. The items related to appetite and weight loaded on the second factor. The third factor included many of the somatic symptoms such as "headaches" and all of the heart related items. Four items do not appear to load on any of the factors and these include: "tired/fatigue," "problems with sleep" and "stomach aches," and "other aches and pains."

Table 6.6

Factor loadings for items on the IDSS-G

	F1	F2	F3
D01 Sad	0.713*	0.107	0.091
D02 no interest	0.688*	-0.001	-0.034
D03 crying	0.579*	0.287*	0.058
D04 hopeless	0.565*	-0.059	0.248*
D05 lonely	0.748*	0.006	-0.062
D06 social withdrawal	0.745*	0.096	-0.077
D07 tired/fatigue	0.282	0.325*	0.359*
D08 weigh too little	0.351	0.731*	-0.013

D09 weigh too much	0.041	-0.594*	0.167
D10 increased appetite	0.070	0.609*	0.275*
D11 sleep problems	0.276*	0.242*	0.278*
D12 trapped	0.903*	-0.019	-0.063
D13 worry	0.692*	-0.055	0.015
D14 worthless	0.565*	-0.001	0.165
D15 headaches	-0.023	0.238*	0.578*
D16 stomach_aches	-0.254	0.298*	0.344*
D17 other_aches	0.198	0.205	0.209
D18 anger	0.549*	-0.199*	0.109
D19 thinking too much	0.784*	-0.197	0.012
D20 confused	0.843*	-0.042	-0.063
D21 heart_weakness	0.067	0.154	0.543*
D22 palpitations	0.079	0.257*	0.600*
D23 heavy_heart	0.009	-0.059	0.910*
D24 heart_pressure	0.112	-0.014	0.861*
D25 heart_pain	-0.033	0.340*	0.550*
D26 psychomotor	0.608*	0.261*	-0.135
D27 concentration	0.605*	0.060	-0.086
D28 imp function	0.598*	0.084	0.144
D29 suicide	0.674*	0.345*	0.050

Pile Sort Activity

The pile sort activity was done by $n = 50$ participants who were presented with all 29 symptoms on the IDSS-G to sort and group together. Participants on average created 5.4 piles (range: 2-14). The title of the most commonly created piles and frequency in which each pile was created is listed in Table 9. The most frequent pile created was related to "feelings" ($n = 16$),

followed by "depression or sadness" (n = 13) and piles related to physical problems (n = 11) and "thinking too much" (n = 11). For participants who only created two piles, these most often were divided between piles for emotions/feelings and piles for physical problems. Nine participants specifically created piles related to "heart problems" which included all symptoms related to the heart in the IDSS-G.

Table 6.9

Title and frequency of each of pile created during the pile sort activity[a]

Pile name	**N**
Feelings	16
Depression/sadness	13
Physical problems	11
Thinking too much	11
Related to disease	11
Heart Problems	10
Past/traumatic events	9
Stress	9
Trust/betrayal	9
Feeling angry	8
Dissatisfaction/disappointment	8
Functioning	6
Confusion	5
Feeling isolated/no one to rely on	5

[a] Piles that fewer than n = 5 people created were not included in the table

Table 10 displays the frequency with which each symptom was placed in the nine most frequently created piles. Most people sorted out symptoms that related to physical illnesses, somatic complaints, and heart issues and created separate piles for these symptoms. Many of the items on the IDSS-G were most frequently placed in the feelings pile, with the exception of the physical symptoms as well as the item: "difficulty doing your usual activities at home or work." The item related to suicide was most commonly placed in the piles related to feelings and the pile titled "thinking too much." The most common symptoms that were placed in the past/traumatic events pile were: "no interest in things" and "moving or speaking so slowly or so fast that others have noticed."

Table 6.10

Frequency of each symptom on the IDSS-G by pile

Symptom	Feelings (n = 16)	Depression (n =13)	Physical Problems (n =11)	Thinking too much (n =11)	Related to disease (n =10)	Heart Problems (n =10)	Past/traumatic events (n =9)	Stress (n =9)	Trust/betrayal (n =9)
D01 Sad	12	4	0	1	0	1	2	2	1
D02 no interest	9	1	0	1	4	0	4	1	0
D03 crying	8	2	3	2	2	0	1	1	3
D04 hopeless	9	4	1	0	0	1	1	2	2
D05 lonely	8	2	1	2	0	0	3	3	3
D06 social withdrawal	6	3	0	0	2	0	2	1	0
D07 tired/fatigue	6	3	6	0	2	0	2	1	0
D08 weigh too little	1	2	9	2	5	0	0	0	0
D09 weigh too much	1	3	8	1	4	0	2	1	0
D10 increased appetite	6	3	3	1	5	1	0	3	0
D11 sleep problems	7	4	5	1	1	1	3	1	2
D12 trapped	10	6	2	0	1	0	1	2	2

D13 worry	9	3	0	0	3	0	2	4	0
D14 worthless	9	2	1	0	0	0	3	2	1
D15 headaches	2	4	8	4	4	0	0	1	0
D16 stomach_aches	1	3	11	0	6	1	0	0	0
D17 other_aches	1	2	8	2	6	1	2	0	0
D18 anger	9	0	3	0	0	1	0	2	1
D19 thinking too much	11	2	0	7	2	0	2	3	0
D20 confused	9	5	1	2	2	0	1	2	1
D21 heart_weakness	4	2	5	1	4	5	1	0	0
D22 palpitations	2	5	8	0	6	6	0	0	0
D23 heavy_heart	1	5	8	0	5	5	0	0	0
D24 heart_pressure	1	4	7	0	6	6	0	0	0
D25 heart_pain	1	3	8	0	6	5	0	0	0
D26 psychomotor	4	0	3	3	0	1	4	1	0
D27 concentration	9	1	3	2	2	1	1	2	3
D29 imp function	3	1	3	0	1	0	2	0	0
D30 suicide	8	3	1	4	2	0	0	2	2

Criterion validity

Table 12 displays average scores for the IDSS-G by diagnostic category as classified by the psychiatrists. Average scores on the IDSS-G were higher among all the different disorder classifications compared to participants classified as not having one of the SCID disorders (n = 63). Paired t-tests indicated statistically significant differences between the mean score on the IDSS-G for participants classified as having any disorder and MDD/dysthymia compared to participants with none of these disorders. There were no statistical differences in mean IDSS-G comparing those with GAD to no disorder and between the classification of either depressive disorder (MDD/Dysthymia) and GAD.

Table 6.12

Average scores on IDSS-G by diagnostic category

	M (SD)	*Range*
Any disorder vs. no disorder		
Any disorder (n = 71)	*0.87* (0.47)*	*0.11-2.44*
No disorder (n = 63)	*0.55* (0.43)*	*0.00-2.11*
MDD/Dysthymia vs. no disorder		
Depression/Dysthymia (n = 52)	*0.93* (0.49)*	*0.11-2.44*
No disorder (n = 63)	*0.55* (0.43)*	*0.00-2.11*
GAD vs. no disorder		
GAD (n = 22)	*0.75* (0.39)	*0.29-1.96*
No disorder (n = 63)	*0.55* (0.43)	*0.00-2.11*
Depression vs. GAD		
Depressive disorders (n = 39)	*1.01* (0.52)	*0.11-2.44*

GAD (n = 18)	*0.72* (0.38)	*0.30-1.96*

* Paired t-tests indicate significant difference ($p < 0.05$) in means between groups

[a] Depression category includes both MDD and Dysthymia.

Incremental Validity

Table 13 presents results from the incremental validity investigation. Based on the construct validity results, four regression models were compared that progressively included all variables found to be associated with scores on the impaired functioning measure.

Model 1 indicated that age was associated with impaired functioning, meaning that for every year increase in age there is a -0.008 point decrease in impaired functioning. Only three percent of the variance in impaired functioning was explained by age. Results from model 2, showed that once the suicidal ideation item was included, there was no more effect for age and that suicidal ideation (dichotomized as any/none) was associated with worsening functioning. Model 2 explained a total of 16% of the variance in impaired functioning. Model 3, included the PHQ-9 which was found to be significantly associated with impaired functioning, after accounting for age and suicidal ideation. For every unit increase in scores on the PHQ-9, there was a 0.37 point increase in impaired functioning. Model 3 explained 28% of the variance in impaired functioning.

The final model (model 4) included all variables from model 3, as well as average scores on the IDSS-G. Thirty three percent of the total variance in impaired functioning was explained by the variables in model 4. Results from model 4 indicate that every unit increase on the IDSS-G was associated with a 0.47 increase in impaired functioning. The IDSS-G significantly predicted worse impaired functioning after controlling for age, suicidal ideation and scores on the PHQ-9. Moreover, after adding the IDSS-G, the PHQ-9 was no longer significantly associated with impaired functioning, thus supporting incremental validity of the IDSS-G to predict functional impairment beyond what is predicted by the PHQ-9.

Table 6.13

Effects of measured variables on impaired functioning presented as beta coefficients

Model	β (SE)	t	Total variance explained by model (R^2)
Model 1			0.03
Age	-0.008 (0.01)	-2.21*	
Model 2			0.16
Age	-0.004 (0.01)	-1.22	
Suicidal ideation[a]	0.71 (0.15)	4.75**	
Model 3			0.28
Age	-0.003 (0.01)	-0.92	
Suicidal ideation	0.25 (0.16)	1.56	
PHQ-9	0.37 (0.08)	4.78**	
Model 4			0.33
Age	-0.003 (0.01)	-0.92	
Suicidal ideation	0.22 (0.16)	1.40	
PHQ-9	0.12 (0.11)	1.21	
IDSS-G	0.47 (0.14)	3.35**	

[a] The item related to suicide ideation was dichotomized meaning that 0 = none of the time and 1 = some, most and almost all of the time.

* $p < 0.05$

** $p < 0.001$

Clinical Utility

In terms of Area Under the Curve (AUC) analysis, the IDSS-G had an AUC of 0.72 for the comparison on of any disorder to no disorder and an AUC of 0.75 when comparing depressive disorders (MDD/Dysthymia) to no disorder. The AUCs for GAD compared to no disorder was

lower (0.68). The IDSS-G could not differentiate accurately between a diagnosis of a depressive disorder (MDD/Dysthymia) and a diagnosis of GAD. For the PHQ-9, AUCs were slightly higher across all diagnostic categories, with the exception of the GAD vs. no disorder comparison (Table 14).

Table 6.14

Area Under the Curves (AUC) for the IDSS-G and PHQ-9 across diagnostic categories

	IDSS-G	*PHQ-9*
	AUC, [95%CI]	*AUC,* [95%CI]
Any disorder vs. no disorder	0.72	0.74
	[0.63, 0.81]	[0.65, 0.82]
MDD/Dysthymia vs. no disorder	0.75	0.76
	[0.65, 0.84]	[0.67, 0.85]
GAD vs. no disorder	0.68	0.68
	[0.56, 0.79]	[0.56, 0.80]
Depression vs. GAD	0.30	0.33
	[0.16, 0.45]	[0.17, 0.50]

Finally, based on Liu et al. (2012), optimal cutoffs for identifying any disorder vs. no disorder and a depressive disorder vs. no disorder were calculated for IDSS-G and PHQ-9. Cutpoints and their corresponding test statistics (sensitivity, specificity, positive predictive value, negative predictive value, percent correctly classified) can be found in Table 15. Using a cutoff of 0.56 on the IDSS-G correctly identified people with any of the disorders 73% of the time and correctly identified individuals as having none of the disorders 67% of the time. Similarly, using

a cutoff value of 0.56 on the IDSS-G correctly identified people with a depressive disorder (MDD/Dysthymia) 77% of the time and correctly identified individuals as having no disorder 67% of the time. For the PHQ-9, using a cutpoint of 0.44 correctly identified those with any of the disorders 82% of the time, and correctly identified individuals without any of the disorders 56% of the time; and correctly identified individuals with a depressive disorder 89% of the time and correctly identified individuals with without any of the disorders 56% of the time.

Table 6.15

Cutoff values for average scores (range: 0-3) and corresponding classification statistics for the IDSS-G and PHQ-9

	IDSS-G Cutpoints			PHQ-9 Cutpoints		
	Optimal (0.56)	High (1.00)	Low (0.3)	Optimal (0.44)	High (1.11)	Low (0.3)
Any disorder vs. no disorder						
Sensitivity	0.73	0.31	0.93	0.82	0.27	0.89
Specificity	0.65	0.83	0.37	0.56	0.89	0.48
Positive Predictive Value	0.70	0.67	0.62	0.67	0.73	0.66
Negative Predictive Value	0.68	0.52	0.82	0.73	0.52	0.79
Correctly classified	0.69	0.55	0.66	0.69	0.56	0.69
MDD/Dysthymia vs. no disorder						
Sensitivity	0.77	0.38	0.94	0.89	0.31	0.92
Specificity	0.65	0.83	0.37	0.56	0.89	0.48
Positive Predictive Value	0.65	0.65	0.55	0.62	0.70	0.59
Negative Predictive Value	0.77	0.62	0.89	0.85	0.61	0.88
Correctly classified	0.70	0.63	0.63	0.74	0.63	0.68

Classification statistics for higher and lower cutoff values were also explored. For the IDSS-G, the high cutoff value was set at 1.00 as this represented an average response of 1 across the 27 items, and the low cutoff value was set at 0.3. At the high cutpoint, sensitivity decreased substantially to 31% for any disorder vs. no disorder and to 38% for depressive disorder vs. no disorder; but specificity increased to 83% for both diagnostic comparisons. For the lower cutpoints, sensitivity increased but specificity decreased. For the PHQ-9, the high cutpoint was set at 1.11, which corresponded to the traditional cutoff of 10 (using score totals rather than averages) found in the literature (Manea, Gilbody, & McMillan, 2012), and the low cutpoint was set at 0.3. Using the high cutpoint, sensitivity decreased to 26.8% for any disorder vs. no disorder and 30.8 for depressive disorder vs. no disorder. At this high cutpoint, specificity increased to 88.9% for both diagnostic comparisons. Using the lower cutpoint increased sensitivity and decreased specificity.

Discussion

The present study examined the reliability and validity of the newly created International Depression Symptom Scale-Global version (IDSS-G), a instrument to measure depression developed based on an empirical investigation into the signs and symptoms of depression that occur in many populations across the world. The IDSS-G was shown to have high internal consistency reliability ($\alpha = 0.92$), test-retest reliability ($r = 0.87$), and inter-rater reliability ($ICC = 0.90$). The PCA, the EFA and the pile sort activity, all suggested a possible 3-factor solution (emotions/social isolation, weight/appetite, somatic complaints), as the most appropriate model.

Cognitive interviewing results indicate that most items on the IDSS-G were understandable and showed a high degree of face validity. Assessment of construct validity showed that the IDSS-G was strongly correlated with the PHQ-9 ($r = 0.78$), impaired functioning (full measure: $r = 0.56$; single item: $Rho = 0.65$), and suicidal ideation ($Rho = 0.66$). Most of the items on the IDSS-G independently predicted worsening impaired function, supporting the

strength/relevance of each individual item on the IDSS-G. In addition, the IDSS-G was shown to be incrementally valid, as higher scores on the IDSS-G predicted worse functioning above and beyond what was predicted by the PHQ-9. The PHQ-9, while widely used and validated in similar contexts, was developed based on western clinical populations. The IDSS-G was developed based on global presentations of depression. Results from the incremental validity analysis suggest that a measure based on a global presentation of depression may be able to predict functional impairment in a non-western population above and beyond a translated, commonly used measure.

Using psychiatrist diagnosis with a semi-structured DSM-IV based instrument, the IDSS-G demonstrated criterion validity. This means that scores on the IDSS-G were significantly higher for people with any of the disorders assessed (MDD, Dysthymia, and/or GAD) compared to people with none of these disorders; and for people with a depressive disorder (MDD and Dysthymia) compared to scores for people with no depressive disorder. As expected, scores on the IDSS-G were not significantly different between people with GAD only compared to people with no disorder, or between people with a depressive disorder compared to people with GAD only.

The IDSS-G had an AUC of 0.72 when used to detect any of the disorders of interest and an AUC of 0.75 when used to detect depressive disorders specifically. This indicates that the IDSS-G has low to moderate diagnostic utility for detecting DSM defined disorders. However, it did perform comparably to the PHQ-9 in this setting, which also showed low to moderate AUCs (any disorder: $AUC = 0.74$; depressive disorder: $AUC = 0.76$ respectively). Both instruments performed worse than western measures in similar studies that compared instrument performance to DSM diagnosis by trained mental health professionals, in non-western settings. Silove and colleagues (2007), found an AUC of 0.83 for the HSCL in a non-trauma affected population in Cambodia. Similarly, Lotrakul et al. (2008) tested the PHQ-9 in a primary health care setting in Bangkok, Thailand and found an AUC of 0.89.

Based on the criterion and clinical utility results the IDSS-G appears to be a better measure for depressive disorders than for general distress in this population. The criterion validity results and the ROC analysis suggest that the IDSS-G can distinguish between combined depression/GAD and no disorder, and depressive disorders and no disorder, but not between GAD alone and no disorder, or depression and GAD. At optimal cutpoints, the IDSS-G had higher sensitivity when detecting people with a depressive disorder (*sens* = 0.77), than a mix of depression and anxiety (*sens* = 0.70). However, distinguishing between depression and anxiety is challenging. There is robust evidence supporting the strong association between depression and anxiety globally (Abas & Broadhead, 1997; Bener, Ghuloum, & Abou-Saleh, 2012). Moreover, depressive disorders and GAD are often comorbid (Kessler & Bromet, 2013), share similar risk factors (Almeida et al. 2012), have similar neurocognitive processes (Ressler & Nemeroff, 2000), respond to similar treatments (Butler, Chapman, Forman, & Beck, 2006); and even share some of the same diagnostic criteria (American Psychological Association, 2015). The IDSS-G's ability to better detect depressive disorders, yet also pick up on some cases of GAD as well, is understandable.

The IDSS-G was found to be uncorrelated with gender, which was surprising given the well-established association of gender and depression in the literature (Culbertson, 1997; A. J. Ferrari et al., 2013; Kuehner, 2003; Nolen-Hoeksema, 1987; Van de Velde, Bracke, & Levecque, 2010). It is important to remember that the sample in this study was not randomly selected and thus, gender differences may not appear because of selection bias. However, another possible explanation is that the IDSS-G was created based on a hypothesis that there are signs and symptoms of depression found across all populations, and includes signs and symptoms frequently reported by both male and female populations. Thus, the lack of association with gender in the current study, but strong associations with other measured variables, may point to the strength of the IDSS-G for detecting depression across genders. This is consistent with a recent study from the United States, that showed when scales incorporate both traditional

symptoms (defined by the DSM) and symptoms that assess "male-type depression" (i.e. externalizing symptoms of depression such as substance abuse, aggression, risk-taking behavior), sex disparities in depression are eliminated (Martin, Neighbors, & Griffith, 2013). The IDSS-G includes the symptom of "anger" which while traditionally thought of as being an expression of "male-type depression" has been found to occur in female populations as well (Rees et al. 2013; Williamson, O'Hara, Stuart, Hart, & Watson, 2014).

Cognitive interviewing indicated that some items, most notably the item related to stomach pains and the items related to heart issues, were considered to be attributable to medical illness. This further supports the results from the EFA and pile sort activities, indicating that items representing somatic complaints are different from the other items on the IDSS-G, and may be unrelated to depression. However, it is important to remember when interpreting these results that all of the participants were recruited from a community medical clinic. It may be that participants are especially aware of medical complaints and therefore view these symptoms as distinct from problems related to emotions and behaviors. Before any revisions to the IDSS-G are made, further studies are needed to determine how the items perform in other populations and settings.

Four items were not independently associated with impaired functioning including "weighing too much," "stomach aches," "other aches and pains" and "pain in the heart." These four items all relate to somatic complaints. Considering the lack of association with impaired functioning and the results from the internal consistency reliability analysis, factor analysis, pile sort, and cognitive interviews, these items appear to be candidates for removal from the IDSS-G. However, further exploration into individual item characteristics should be done before any items are removed, but this type of analysis is beyond the scope of the current study.

Researchers deciding between the use of the IDSS-G or the PHQ-9 for future studies in this population, should take into account the incremental validity and diagnostic classification results, as well as practical considerations. If the goal of a future study is to accurately identify

people with DSM defined MDD/dysthymia, then the PHQ-9, given its higher sensitivity, may be a better instrument for use in this population. The PHQ-9 is also a shorter instrument, making it easier to administer in settings where time is limited (such as screening in primary health care). If the goal of a study is to screen for people with a depressive disorder, who are also suffering from depression symptoms that impair functioning, than the IDSS-G may be the preferable tool. The IDSS-G compared to the PHQ-9, showed slightly lower sensitivity and higher specificity, across diagnostic comparisons. However, the IDSS-G was able to better predict impaired functioning than the PHQ-9. Taken together the evidence suggests that the IDSS-G does a moderate job of screening people for DSM defined depressive disorders and can accurately identify people with impaired functioning. However, the IDSS-G is quite long and further work needs to be done to investigate the impact of shortening the measure on its psychometric properties.

The promising reliability and validity results for the IDSS-G provide preliminary evidence to support its use as a screening tool for depressive disorders in low-income settings. Of note, the IDSS-G was translated and back translated, but did not undergo preliminary cultural adaptation. Rather its development was based on empirical evidence related to global presentations of depression. Despite agreement that simple translation and back-translation is not sufficient when using mental health measures in other settings (Kohrt et al., 2011) and many studies showing that adapted western-based instruments can be reliable and valid in other contexts (Bass et al., 2008a; Betancourt et al., 2009; Bolton, 2001; Haroz et al., 2014a; Rasmussen et al., 2014), there have been few studies which have actually evaluated the impact of cultural adaptation on scale validity. Jayawickreme et al. (2012) conducted a study looking at the incremental validity of western psychological measurement instruments that incorporated local idioms of distress. Base on their results, the authors concluded that instruments that incorporate local idioms of distress predicted functional impairment above and beyond simple translations of well-established western measures. Based on this conclusion, the authors stress the importance of doing brief ethnographic work to inform scale adaptation and point to the DIME approach by

Bolton and colleagues (Applied Mental Health Research Group, 2013) as an example of this type of qualitative investigation. The results from the present study suggest that it may also be possible to create a measurement instrument based on global presentations of depression, which could be used to accurately screen for depression when preliminary qualitative work is not feasible. The IDSS-G, or similar measure, may serve as better instruments than translated versions of well-established western-based measures.

However, the development of the IDSS-G does not diminish the importance of locally relevant signs and symptoms of distress. When identified, these local indicators are important, if not the most important indicators to include on measures in each context, as they represent common ways of expressing distress in each setting (Keys, Kaiser, Kohrt, Khoury, & Brewster, 2012; Kohrt et al., 2014). Local expressions of distress may more saliently communicate illness, may be less stigmatizing, and may be useful for measuring treatment success (Kohrt et al., 2014). One way to address this concern was by explicitly creating the IDSS with the intention of being flexible. The IDSS includes a global measure of symptoms thought to be related to depression in many contexts (IDSS-G), and a local measure that augments the global measure with locally relevant symptoms (IDSS-L). Results from this study show that the IDSS-G performed well, and the addition of local symptoms (IDSS-L) did not markedly change the performance of the measure in this context (see Appendix F).

Given the dearth of resources available to systematically identify common signs and symptoms of depression, adapt instruments and then test these instruments in the local context, the IDSS-G has the benefit of being informed by depressive illness' symptom presentation in LMIC and could be a better starting place if resources are limited or unavailable. Moreover, IDSS-G was informed by IRT methodology, which allows for estimating sample independent parameters of the items and provides a powerful framework to identify potential bias and account for this bias when comparing scores.

Limitations

The present study has a number of limitations. First, the study sample involved a non-random sample in an urban setting who had medical illness. Findings from this study may not generalize to populations outside of Yangon, Myanmar. Second, many of the symptoms included in the IDSS-G are based on English translations of signs and symptoms found in qualitative research focused on depression. However, language is nuanced, and a direct translation can at times not accurately capture the full meaning of a phrase. In fact, most research has indicated that despite efforts to find accurate translations of symptoms and syndromes from one language to another, direct translation often results in overlapping terms that do not necessarily fully capture the meaning of the original term (Nichter, 2010). Thus, it may be that some symptoms on the translated Myanmar version of the IDSS-G may not full capture how distress is conveyed locally.

All of the psychiatrists who provided criterion ratings had been medically trained and currently practice in Yangon, Myanmar, but none of them had experience using a standardized clinical interview before this study. As a result misclassification of participants may have been possible. However, given the inter-rater reliability results this is rather unlikely. This highlights another challenge for validity studies in LMIC: even when it is possible to find mental health professionals to participate in a validity study, finding ones trained in using standardize questionnaires, is challenging. There is an ongoing need for effective methods for validation of measurement instruments in places without these resources.

Another limitation relates to the use of DSM-IV diagnoses as the criterion. This use of DSM diagnoses as the criterion (or gold standard) has been described as a reification problem in the literature (Hyman, 2010). The DSM consists of disorders that are meant to be heuristics, but due to various clinical, policy, and legal forces, these disorders have become reified (Hyman, 2010). DSM diagnoses, while currently defined as a gold standard, may not be the most accurate definition of mental disease. In the current study, the focus on DSM ratings may have overlooked local cultural concepts of distress that do not necessarily fit into western psychiatric nosology but

are particularly important (Kohrt et al. 2014). For example, results from the pile sort activity showed that participants grouped some of the symptoms as being related to *thinking too much*. *Thinking too much* may be its own illness category, which causes significant distress and impairs functioning, and would be perhaps a better criterion in this setting. However, using this as a criterion was not possible and the results from the current study only support the criterion validity of the IDSS-G for detecting DSM defined depressive disorders.

Finally, items on the IDSS-G were also partially based on the HSCL 15-item depression scale (Mollica et al., 2004; Winokur et al., 1984) because that was the basis for the quantitative IRT analyses that informed the IDSS-G development. Thirteen of the items on the HSCL 15-item depression scale are included in the IDSS-G. In a previous study in a similar population, the HSCL-15 and an additional item related to disappointment (included on the IDSS-L) was shown to be reliable and valid (Haroz et al., 2014a) Thus, it is possible that the reason the IDSS-G better predicted functional impairment when compared with the PHQ-9, was because it was largely based on a measure that had been informed by qualitative data and adapted for use in a similar population. However, no single item better predicted functional impairment, suggesting that all 27 items on the IDSS-G were important for measuring depression in this context and that the measure as a whole was incrementally more valid than the PHQ-9.

Conclusion

Overall, the findings show that the IDSS-G is a reliable and valid measurement instrument for depression for use in Yangon, Myanmar. Results support the use of the IDSS-G as a screening measure to detect depression rather than general distress in this context. Incremental validity results suggest that the IDSS-G, which is based on global presentations of depression, may be a better measure for depression in low-resource, non-western contexts, than the use of a translated measure developed for western populations. Thus, in situations where resources for formative qualitative and quantitative work are limited, use of the IDSS-G seems particularly

practical. However, the importance of local idioms of distress should not be minimized and when possible incorporated into measurement and screening. These results, taken together, provide preliminary evidence to support the ongoing testing and refinement of the IDSS-G for use by lay-workers in LMIC for screening and intervention monitoring purposes. Further work needs to be done to potentially reduce the number of items on the IDSS-G and investigate its use as a guide for clinical decisions and cross-cultural epidemiological work.

References

Abas, M. A., & Broadhead, J. C. (1997). Depression and anxiety among women in an urban setting in Zimbabwe. *Psychological Medicine, 27*(01), 59-71.

Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *Journal of Affective Disorders, 96*(1-2), 89-93.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Almeida, O. P., Draper, B., Pirkis, J., Snowdon, J., Lautenschlager, N. T., Byrne, G., . . . Flicker, L. (2012). Anxiety, depression, and comorbid anxiety and depression: Risk factors and outcome over two years. *International Psychogeriatrics, 24*(10), 1622-1632.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR.* Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders fifth edition* (5th ed.*)*. Washington, DC: Author.

Applied Mental Health Research Group (AMHR). (2013). *Design, implementation, monitoring and evaluation of cross-cultural trauma related mental health and psychosocial assistance programs: A user's manual for researchers and program implementers*. Retrieved from: http://www.jhsph.edu/research/centers-and-institutes/center-for-refugee-and-disaster-response/response_service/AMHR/dime

Bass, J. K., Annan, J., McIvor Murray, S., Kaysen, D., Griffiths, S., Cetinoglu, T., . . . Bolton, P. A. (2013). Controlled trial of psychotherapy for Congolese survivors of sexual violence. *New England Journal of Medicine, 368*(23), 2182-2191.

Bass, J. K., Ryder, R. W., Lammers, M., Mukaba, T. N., & Bolton, P. A. (2008). Post-partum depression in Kinshasa, democratic republic of Congo: Validation of a concept using a mixed-methods cross-cultural approach. *Tropical Medicine & International Health, 13*(12), 1534-1542.

Bener, A., Ghuloum, S., & Abou-Saleh, M. T. (2012). Prevalence, symptom patterns and comorbidity of anxiety and depressive disorders in primary care in Qatar. *Social Psychiatry and Psychiatric Epidemiology, 47*(3), 439-446.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.* 289-300.

Betancourt, T. S., Bass, J., Borisova, I., Neugebauer, R., Speelman, L., Onyango, G., & Bolton, P. (2009). Assessing local instrument reliability and validity: A field-based example from northern Uganda. *Social Psychiatry and Psychiatric Epidemiology, 44*(8), 685-692.

Bolton, P., Bass, J., Neugebauer, R., Verdeli, H., Clougherty, K. F., Wickramaratne, P., . . . Weissman, M. (2003). Group interpersonal psychotherapy for depression in rural uganda: A randomized controlled trial. *JAMA, 289*(23), 3117-3124.

Bolton, P., Lee, C., Haroz, E. E., Murray, L., Dorsey, S., Robinson, C., . . . Bass, J. (2014). A transdiagnostic community-based mental health treatment for comorbid disorders: Development and outcomes of a randomized controlled trial among Burmese refugees in Thailand. *PLoS Medicine, 11*(11), e1001757.

Bolton, P. (2001). Cross-cultural validity and reliability testing of a standard psychiatric assessment instrument without a gold standard. *The Journal of Nervous and Mental Disease, 189*(4), 238-242.

Bromet, E., Andrade, L. H., Hwang, I., Sampson, N. A., Alonso, J., de Girolamo, G., . . . Kessler, R. C. (2011). Cross-national epidemiology of DSM-IV major depressive episode. *BMC Medicine, 9*(90).

Bruckner, T. A., Scheffler, R. M., Shen, G., Yoon, J., Chisholm, D., Morris, J., . . . Shekhar, S. (2011). The mental health workforce gap in low- and middle-income countries: A needs-based approach. *Bulletin of the World Health Organization, 89*(3), 184-194.

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26*(1), 17-31.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Culbertson, F. M. (1997). Depression and gender: An international review. *American Psychologist, 52*(1), 25-31.

Ertl, V., Pfeiffer, A., Saile, R., Schauer, E., Elbert, T., & Neuner, F. (2010). Validation of a mental health assessment in an African conflict population. *Psychological Assessment, 22*(2), 318-324.

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J. L., . . . Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010. *PLoS Medicine, 10*(11), e1001547.

Ferrari, A., Somerville, A., Baxter, A., Norman, R., Patten, S., Vos, T., & Whiteford, H. (2012). Global variation in the prevalence and incidence of major depressive disorder: A systematic review of the epidemiological literature. *Psychological Medicine, 43*(3), 1-11.

First, M.B., Spitzer, R.L., Gibbon, M. & Williams, J.B.W. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP).* New York: Biometrics Research, New York State Psychiatric Institute, November 2002.

Fischer, J. E., Bachmann, L. M., & Jaeschke, R. (2003). A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine, 29*(7), 1043-1051.

Fleiss, J. (1986). *The Design and Analysis of Clinical Experiments*. New York, NY: John Wiley & Sons.

Ghimire, D. J., Chardoul, S., Kessler, R. C., Axinn, W. G., & Adhikari, B. P. (2013). Modifying and validating the composite international diagnostic interview (CIDI) for use in Nepal. *International Journal of Methods in Psychiatric Research, 22*(1), 71-81.

Goldberg, D. P., & Huxley, P. (1992). *Common mental disorders: A bio-social model.* New York, NY: Routledge.

Haroz, E. E., Bass, J. K., Lee, C., Murray, L. K., Robinson, C., & Bolton, P. (2014). Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the Thailand Burma border. *BMC Psychology, 2*(1), 31-40.

Hesbacher, P. T. (1980). Psychiatric illness in family practice. *Journal of Clinical Psychiatry; Journal of Clinical Psychiatry, 41,* 6-10.

Hollifield, M. (2002). Accurate measure in cultural psychiatry: Will we pay the costs? *Transcultural Psychiatry, 39*, 419–421.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology, 6*, 155-179.

Jayawickreme, N., Jayawickreme, E., Atanasov, P., Goonasekera, M., & Foa, E. B. (2012). Are culturally specific measures of trauma-related anxiety and depression needed? The case of Sri Lanka. *Psychological Assessment, 24*(4), 791-800.

Jorm, A. F. (2000). Does old age reduce the risk of anxiety and depression? A review of epidemiological studies across the adult life span. *Psychological Medicine, 30*(01), 11-22.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., . . . Wang, P. S. (2003). The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (NCS-R). *JAMA, 289*(23), 3095-3105.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual Review of Public Health, 34*(1), 119-138.

Keys, H. M., Kaiser, B. N., Kohrt, B. A., Khoury, N. M., & Brewster, A. T. (2012). Idioms of distress, ethnopsychology, and the clinical encounter in Haiti's central plateau. *Social Science & Medicine, 75*(3), 555-564.

Kohrt, B. A., Jordans, M. J. D., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). Validation of cross-cultural child mental health and psychosocial research

instruments: Adapting the depression self-rating scale and child PTSD symptom scale in nepal. *BMC Psychiatry, 11-27.*

Kohrt, B. A., Rasmussen, A., Kaiser, B. N., Haroz, E. E., Maharjan, S. M., Mutamba, B. B., . . . Hinton, D. E. (2014). Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations for global mental health epidemiology. *International Journal of Epidemiology, 43*(2), 365-406.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*(9), 1-7.

Kuehner, C. (2003). Gender differences in unipolar depression: An update of epidemiological findings and possible explanations. *Acta Psychiatrica Scandinavica, 108*(3), 163-174.

Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in Medicine, 31*(23), 2676-2686.

Lotrakul, M., Sumrithe, S., & Saipanish, R. (2008). Reliability and validity of the thai version of the PHQ-9. *BMC Psychiatry, 8*, 46-53.

Lund, C., Tomlinson, M., De Silva, M., Fekadu, A., Shidhaye, R., Jordans, M., . . . Prince, M. (2012). PRIME: A programme to reduce the treatment gap for mental disorders in five low- and middle-income countries. *PLoS Medicine, 9*(12), e1001359.

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the patient health questionnaire (PHQ-9): A meta-analysis. *CMAJ: Canadian Medical Association Journal, 184*(3), E191-6.

Marc, L. G., Henderson, W. R., Desrosiers, A., Testa, M. A., Jean, S. E., & Akom, E. E. (2014). Reliability and validity of the haitian creole PHQ-9. *Journal of General Internal Medicine, 29*(12), 1679-1686.

Martin, L. A., Neighbors, H. W., & Griffith, D. M. (2013). The experience of symptoms of depression in men vs women: Analysis of the national comorbidity survey replication. *JAMA Psychiatry, 70*(10), 1100-1106.

Mels, C., Derluyn, I., Broekaert, E., & Rosseel, Y. (2010}). Community-based cross-cultural adaptation of mental health measures in emergency settings: Validating the IES-R and HSCL-37A in eastern democratic republic of Congo. *Social Psychiatry and Psychiatric Epidemiology, 45*(9), 899-910.

Miller, K. E., Omidian, P., Quraishy, A. S., Quraishy, N., Nasiry, M. N., Nasiry, S., . . . Yaqubi, A. A. (2006). The Afghan symptom checklist: A culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry, 76*(4), 423-433.

Mollica, R. F., McDonald, L. S., Massagli, M. P., & Silove, D. M. (2004). *Measuring trauma, measuring torture.* Cambridge, MA: Harvard University.

Mulrow, C. D., Williams, J. W., Gerety, M. B., Ramirez, G., Montiel, O. M., & Kerber, C. (1995). Case-finding instruments for depression in primary care settings. *Annals of Internal Medicine, 122*(12), 913-921.

Murray, C. J., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., . . . Abdalla, S. (2013). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010. *The Lancet, 380*(9859), 2197-2223.

Murray, L. K., Dorsey, S., Haroz, E., Lee, C., Alsiary, M. M., Haydary, A., . . . Bolton, P. (2013). A common elements treatment approach for adult mental health problems in low-and middle-income countries. *Cognitive and Behavioral Practice, 21*(2), 111-123.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (Seventh Edition ed.). Los Angeles, CA: Muthén & Muthén.

Nichter, M. (2010). Idioms of distress revisited. *Culture, Medicine, and Psychiatry, 34*(2), 401-416.

Nock, M. K., Borges, G., Bromet, E. J., Alonso, J., Angermeyer, M., Beautrais, A., . . . Williams, D. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry : The Journal of Mental Science, 192*(2), 98-105.

Nolen-Hoeksema, S. (1987). Sex differences in unipolar depression: Evidence and theory. *Psychological Bulletin, 101*(2), 259-282.

Nolen-Hoeksema, S., Larson, J., & Grayson, C. (1999). Explaining the gender difference in depressive symptoms. *Journal of Personality and Social Psychology, 77*(5), 1061-1072.

Ormel, J., Petukhova, M., Chatterji, S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., . . . Kessler, R. C. (2008). Disability and treatment of specific mental and physical disorders across the world. *The British Journal of Psychiatry : The Journal of Mental Science, 192*(5), 368-375.

Patel, V., Simon, G., Chowdhary, N., Kaaya, S., & Araya, R. (2009). Packages of care for depression in low-and middle-income countries. *PLoS Medicine, 6*(10), e1000159.

Patel, V., Simunyu, E., Gwanzura, F., Lewis, G., & Mann, A. (1997). The Shona symptom questionnaire: The development of an indigenous measure of common mental disorders in harare. *Acta Psychiatrica Scandinavica, 95*(6), 469-475.

Patel, V., Araya, R., Chowdhary, N., King, M., Kirkwood, B., Nayak, S., . . . Weiss, H. (2008). Detecting common mental disorders in primary care in India: A comparison of five screening questionnaires. *Psychological Medicine, 38*(02), 221-228.

Patel, V., Chowdhary, N., Rahman, A., & Verdeli, H. (2011). Improving access to psychological treatments: lessons from developing countries. *Behaviour research and therapy*, *49*(9), 523-528.

Patel, V. (2007). *Mental health in low- and middle-income countries.* British Medical Bulletin, 81-82.

Phan, T., Steel, Z., & Silove, D. (2004). An ethnographically derived measure of anxiety, depression and somatization: The Phan Vietnamese psychiatric scale. *Transcultural Psychiatry, 41*(2), 200-232.

Rahman, A., Malik, A., Sikander, S., Roberts, C., & Creed, F. (2008). Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: A cluster-randomised controlled trial. *Lancet, 372*(9642), 902-909.

Rasmussen, A., Eustache, E., Raviola, G., Kaiser, B., Grelotti, D. J., & Belkin, G. S. (2014). Development and validation of a haitian creole screening instrument for depression. *Transcultural Psychiatry, 52*(1), 33-57.

Rees, S., Silove, D., Verdial, T., Tam, N., Savio, E., Fonseca, Z., . . . Tay, K. (2013). Intermittent explosive disorder amongst women in conflict affected Timor-Leste: Associations with human rights trauma, ongoing violence, poverty, and injustice. *PloS One, 8*(8), e69207.

Ressler, K. J., & Nemeroff, C. B. (2000). Role of serotonergic and noradrenergic systems in the pathophysiology of depression and anxiety disorders. *Depression and Anxiety, 12*(S1), 2-19.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annu.Rev.Psychol., 59*, 419-450.

Silove, D., Manicavasagar, V., Mollica, R., Thai, M., Khiek, D., Lavelle, J., & Tor, S. (2007). Screening for depression and PTSD in a Cambodian population unaffected by war. *The Journal of Nervous and Mental Disease, 195*, 152-157.

StataCorp. (2013). *Stata statistical software* (Release 13 ed.). College Station, TX: StataCorp LP.

Van de Velde, S., Bracke, P., & Levecque, K. (2010). Gender differences in depression in 23 european countries. cross-national variation in the gender gap in depression. *Social Science & Medicine, 71*(2), 305-313.

van Ginneken, N., Tharyan, P., Lewin, S., Rao, G. N., Meera, S., Pian, J., . . . Patel, V. (2013). Non-specialist health worker interventions for the care of mental, neurological and substance-abuse disorders in low-and middle-income countries. *The Cochrane Database of Systematic Reviews, 11*, CD009149-CD009149.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Fam Med, 37*(5), 360-363.

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., . . . Gureje, O. (2007). Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *The Lancet, 370*(9590), 841-850.

Williamson, J. A., O'Hara, M. W., Stuart, S., Hart, K. J., & Watson, D. (2015). Assessment of postpartum depressive symptoms: The importance of somatic symptoms and irritability. *Assessment*, *22*(3), 309-318.

Winokur, A., Winokur, D. F., Rickels, K., & Cox, D. S. (1984). Symptoms of emotional distress in a family planning service: Stability over a four-week period. *The British Journal of Psychiatry, 144*(4), 395-399.

World Health Organization. (2011). *Mental health atlas, 2011.* Geneva: World Health Organization.

Zhong, Q., Gelaye, B., Fann, J. R., Sanchez, S. E., & Williams, M. A. (2014). Cross-cultural validity of the spanish version of PHQ-9 among pregnant peruvian women: A rasch item response theory analysis. *Journal of Affective Disorders, 158*, 148-153.

Chapter 7. Discussion

This thesis was designed to explore: 1) how depression is expressed in populations all over the world; and 2) if existing measurement instruments adequately capture the presentation of depression in non-western contexts, or if there is a need to develop a new instrument to do so. Methods included identification of signs and symptoms of depression found in qualitative literature, evaluation of the performance of items from a commonly used measure of depression across diverse settings, and examination of the psychometric properties of the International Depression Symptom Scale (IDSS), a newly created measurement instrument for depression.

Accurate measurement of depression is important for psychiatric epidemiology, clinical screening, program monitoring and evaluation, and clinical decision making purposes. Existing measures of depression are based on Western-clinical presentations of depression, making their broad applicability questionable. Adaptation of these instruments for use in non-western settings can be resource intensive. To address these limitations this study sought to answer the following questions:

1. What are the signs and symptoms of depression that are most frequently reported in qualitative literature from a range of cultures and settings?
2. Are there particular signs and symptoms of depression that are included in western-based measures that are applicable and unbiased across settings?
3. Is it possible to develop a reliable and valid measure of depression that is based on the signs and symptoms of depression that present globally?

The main findings related to these questions were summarized in depth in each of the following chapters: Chapter 4, *Global signs and symptoms of depression: A systematic review of the qualitative literature*; Chapter 5, *Depression symptoms across settings: An IRT analysis of the Hopkins Symptom Checklist for depression using data from eight diverse studies*; and Chapter 6, *Development, reliability, and validity of the International Depression Symptom Scale (IDSS): A measurement instrument for global presentations of depression.*

7.1 Summary of findings

The major findings from this study demonstrated that there is a group of symptoms related to depression that is common in many populations all over the world, as well as other symptoms that are more informative in certain settings compared to other settings. More specifically, results indicated that: 1) DSM-V symptoms of depression are common and occur globally; 2) most symptoms outside the DSM, but which are included in existing western-based measurement instruments, also perform well; 3) there are additional symptoms that are common across populations, but are not part of current diagnostic criteria or commonly used measurement instruments; 4) different symptoms of depression are more informative in different settings; and 5) a measurement instrument created to more accurately reflect the core global presentation of depression, and with flexibility to incorporate locally salient symptoms, shows promising results in one setting, despite limited formative research to inform its use in the local context.

First, a major finding from both the systematic literature review and IRT analysis was that the signs and symptoms of depression included in DSM-V diagnostic criteria for Major Depressive Disorder (MDD) (American Psychiatric Association, 2013) occur globally. Six out of the top ten most frequently mentioned symptoms of depression globally are also part of diagnostic criteria for MDD. DSM symptoms represented the most frequently mentioned symptom in every region with the exception of Southeast Asia. In the quantitative IRT analysis all of the symptoms from the DSM-V diagnostic criteria (8 out of the 15 items on the HSCL), with the exception of the suicidal ideation item, performed well across settings (high discrimination parameters, relatively infrequent DIF). This further supports the relevance of the symptoms included in DSM-V diagnostic criteria for MDD in non-western settings and is consistent with recent research in the United States suggesting that individuals respond similarly to symptoms included in DSM diagnostic criteria for MDD regardless of race, ethnicity or gender (Uebelacker, Strong, Weinstock, & Miller, 2009).

However, some symptoms in current DSM criteria were shown to not be universally applicable. The DSM criterion related to *psychomotor agitation or slowing* was rarely mentioned across study populations and performed questionably in the instrument testing study. The results from the cognitive interviews showed that participants did not understand the meaning of this item as it relates to depression. This symptom can be challenging to evaluate, particularly in self-report scales, as it is the only diagnostic criterion, which is not purely subjective in nature.

The symptom related to *suicide ideation* was also problematic. Although commonly reported in studies included in the qualitative review, this item performed poorly in the quantitative IRT analysis. This is consistent with a recent review suggesting that suicidal ideation in LMIC may be more related to impulse control disorders, than to mood disorders (Nock et al., 2008). In the current study, other unmeasured factors may have been driving responses to this item, resulting poor performance. If this is the case, than inclusion of this item in summary scores is problematic. However, given the ethical responsibility of performing a suicide risk assessment in any mental health clinical or research setting, it is important to still evaluate suicidal risk in some capacity during screenings.

The second major finding suggests that there are additional symptoms that arose frequently across all regions, quantitatively performed well across settings, but which are currently not part of DSM-V diagnostic criteria for MDD. These include feelings of *social isolation* and *hopelessness*. These symptoms were common across all study populations included in the literature review and performed well across all eight settings included in the quantitative analysis.

Third, some symptoms that were not captured by current diagnostic criteria or commonly used measurement instruments, but appeared to be ubiquitous in global presentations of depression. These include: *anger*, *thinking too much*, and *somatic complaints*. Anger has most commonly been associated with expression of depression only in men (Brownhill, Wilhelm, Barclay, & Schmied, 2005; Cochran & Rabinowitz, 2003), but results from the current study

suggest that it is, in fact, a common symptom of depression for both men and women. *Thinking too much* often arises in qualitative studies as a common way to express mental distress and, at least in the context of PTSD, has been found to be a better predictor of clinically significant distress than DSM based symptoms (Hinton, Reis, & Jong, 2015). Results from the current study show that *thinking too much* is also a common symptom of depression globally, and it predicts impaired functioning in Burmese adults.

The findings related to somatic complaints were mixed. The literature review and quantitative IRT analysis results support the universality of somatic complaints. Somatic symptoms arose frequently across study populations and generally performed well across settings. However, results from the instrument testing study indicated that somatic complaints were not related to depression in the local context. Previous literature has found that somatic complaints are near universal (Draguns & Tanaka-Matsumi, 2003; Uebelacker et al., 2009) and may be particularly common when individuals do not have access to primary health care (Simon, Gater, Kisely, & Piccinelli, 1996). In the sample for the instrument testing project, all participants were existing patients at primary health care clinics. Participants may not have made the connection between somatic symptoms and depression because they were already receiving adequate medical care. Despite the universality of somatic complaints linked to depression, broader societal influences, may impact the importance of these symptoms in evaluating depression in particular contexts.

Forth, evidence from the current study suggests that different symptoms were more informative of depression in different settings. For example, results from the quantitative analysis indicated that *crying a lot* was strongly associated with depression in all settings, except for Dohuk. Similarly, in Thailand, participants had to have higher severity of depression to endorse *appetite problems*, compared to participants in all other settings, indicating that this item is more informative of more severe depression in Thailand and less severe depression in other settings. This variability does not undermine the relevance of the core symptoms of global depression, but

rather speaks to the necessity to account for potential response bias when comparing scores on depression measures across settings.

Finally, the International Depression Symptom Scale (IDSS), a measurement instrument created to reflect global presentations of depression, did not undergo formative qualitative work to inform its adaptation to the local setting, prior to testing in Yangon, Myanmar. Results from the instrument testing study, showed that when compared to the Patient Health Questionnaire (PHQ-9; Kroenke & Spitzer, 2002) the IDSS had similar clinical utility, but predicted worse impaired functioning. This supports preliminary evidence for the incremental validity of the IDSS over a commonly used western-based measure for use in adult populations in Yangon, Myanmar.

7.2 Implications for public health research

The findings of this thesis have several implications for public health, practice and research. The results from this study provide a more comprehensive picture of how depression presents globally. The results from both Aim 1 and Aim 2 suggest a core group of symptoms that is mostly unbiased across settings. Many of the symptoms in this core group are already included in the DSM-V and existing measures of depression. From an epidemiologic perspective, these findings suggest that much of the heterogeneity in cross-national prevalence estimates cannot be attributed to the validity of the concept of depression across settings or to measurement factors. This is consistent with previous research that found symptom patterns of depression to be similar across 15 different countries (Simon, Goldberg, Von Korff, & Üstün, 2002) and a recent review of the epidemiology of depression across cultures, which concludes that discrepancies in cross-national prevalence estimates are not due to methodological factors (Kessler & Bromet, 2013).

However, in the context of trauma, the picture is less clear. Depression prevalence estimates among populations affected by armed conflict or displacement have been found to range from 3% to 85.5% (Steel et al., 2009). An estimated, 27.7% of the variance in these estimates can be attributed to methodological factors (Steel et al. 2009). Based on the data

analyzed as part of the quantitative IRT analysis, the evidence from the current study demonstrates that item level response bias can have a significant impact on aggregate measures of depression. While, these findings, were not based on random samples, they suggest a possible source of bias that should be explored and accounted for in future studies with trauma-affected populations.

This was the first systematic review of emic literature related to depression, which is one of the most common mental disorders globally and among the leading causes of disability worldwide. Only, one other study has used similar methodology to extensively examine symptoms of PTSD in emergency settings outside of North America and Europe (Rasmussen, Keatley, & Joscelyne, 2014). With increasing calls for use of qualitative methods (Bass, Bolton, & Murray, 2007; De Jong & Van Ommeren, 2002) to inform global mental health related research, the systematic review methodology has the potential so synthesize emic research across settings, in order to inform conceptualization and measurement of various mental disorders globally.

The use of Item Response Theory (IRT) as the basis for scale development has important research implications as well. Item Response Theory has several properties that make it particularly useful for scale development in the context of global mental health. This method allows for evaluation of item characteristics independently from the sample to which it was administered (Hambleton, Waminathan, & Rogers, 1991). In the current study, IRT provided the methodology to evaluate which items were most strongly related to depression, where along the depression severity continuum were items most informative and reliable, and whether these properties differed by setting. As few existing measures of depression have been developed based on IRT methods, the present study demonstrates the utility of this method to identify the best performing items within and across settings and account for potential response bias.

7.3 Implications for Public Health Practice

Beyond epidemiology, clearly defining how depression presents globally has implications for more accurate measuring of depression for clinical and research purposes. Including symptoms that are commonly reported globally, but not currently captured in commonly used measures of depression, can improve sensitivity of measurement tools for screening, monitoring and evaluation.

A major challenge in the field of global mental health is how to accurately assess individuals in need of services and monitor the effectiveness of those services, without the availability of trained mental health professionals. As there are few trained mental health professionals in many LMIC (World Health Organization, 2011), comprehensive clinical assessments are a near impossibility in most situations. Screening and monitoring tools that are easy to administer, freely available, and show demonstrated reliability and validity, are incredibly important in these settings.

Generally, research to date has demonstrated that culturally adapted western-based self-report measures can be used by non-professionals to effectively screen for mental health problems and guide treatment decisions in LMIC (Bass, Ryder, Lammers, Mukaba, & Bolton, 2008; Bolton et al., 2014; Haroz et al., 2014; Kohrt et al., 2011). However, adaptation can be resource intensive (Hollifield, 2002), making it a process that individuals or organizations are unlikely to undertake if resources are scarce. In these situations, non-adapted, western-measures are often simply translated and used without locally demonstrated reliability and validity (Bass, Bolton, & Murray, 2007; Kohrt et al., 2011). Results from the current study provide initial evidence that a measurement instrument based on global presentations of depression (e.g. the IDSS) is reliable and valid in one low-resource setting despite the absence of initial cultural adaptation.

The use of IRT methods also has implications for public health practice. IRT methods serve as the basis for computer adaptive testing. Measurement of depression using computer

adaptive measurement instruments would potentially make screening more efficient and reduce respondent burden. In the United States, the National Institute of Health has initiated the Patient-Reported Outcomes Measurement Information System (PROMIS®) which consists of computer adapted measurement instruments aimed at assessing emotional distress, pain, fatigue, sleep disturbance, physical functioning and social participation (for detailed information, see www.nighpromis.org). Utilizing the results from the current study, a similar initiative could be undertaken for global mental health.

7.4 Limitations

When interpreting the results of this study, there are a number of limitations that should be considered. A major limitation of the overall study is the dependence on translated language. Items included in the IDSS were based on translated versions of symptoms that arose in the literature review or were included in the HSCL. Research has shown that despite best efforts to find accurate translations from one language to another, translation often results in overlapping terms that do not necessarily fully capture the meaning of the term in the original language (Nichter, 2010). It is possible, that the items on the IDSS do not fully reflect the meaning of depression symptoms in every context.

Another limitation is the reliance on non-random and/or samples from trauma-affected populations. Only studies involving non-random samples were included in the review of qualitative literature. As such the symptoms mentioned in each of the studies may be only relevant in the specific study context. There may also be other symptoms that represent common ways of expressing depression that did not arise in the literature because they were not common in the specific study sample. The quantitative IRT analysis only included studies done with non-random, trauma-affected populations, thereby limiting the generalizability of the findings. Since the IDSS was developed based on these findings, and was then tested in a non-random sample, results from the instrument testing project have limited external validity as well. However, it is

promising that despite these limitations in external validity, the IDSS still performed well when tested in a new context with a non-trauma affected study population.

Finally, another overall limitation to the current study is the focus on depression as defined by the DSM. Just because symptoms can be identified in a variety of settings does not mean they have the same meaning in all settings (a phenomenon described as category fallacy (Kleinman, 1977). All of the articles included in the qualitative literature review were in English, suggesting that much of this research was conducted by western researchers. Western researchers seeking to better understand depression in other cultures may have explicit and implicit biases. This may cause them to recognize symptoms that are consistent with western psychiatric nosology and attribute these symptoms to western definitions of depression. In addition, the IDSS was tested using a criterion of DSM defined depressive disorders. DSM based diagnoses may not be the most accurate definitions of mental disorder in every setting.

7.4 Next Steps

To address some of the limitations specified here, future work should be done to refine the IDSS and explore its utility for different purposes. The items on the IDSS which were included based on the review of qualitative literature, have not yet undergone the same scrutiny as the items that were analyzed as part of the quantitative IRT analysis. Using IRT to analyze the data collected from the IDSS testing project would be useful for scale shortening and refinement. Replication of the reliability and validity results in other settings and involving randomly selected populations would provide further support for the utility of the IDSS in various settings. Future studies should explicitly investigate the IDSS' use as a clinical monitoring tool to measure response to treatment, as this was not possible in the cross-sectional instrument testing study.

The research presented here offers an example of a novel and useful approach to studying symptoms of mental disorders across populations. By combining systematic review methodology and IRT, this study was able to utilize previously collected data to examine which, if any signs

and symptoms of depression are universal, and to use this information as the basis for development of a new measurement instrument, the International Depression Symptom Scale.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders fifth edition* (5th ed.*)*. Washington, DC: Author.

Bass, J. K., Bolton, P. A., & Murray, L. K. (2007). Do not forget culture when studying mental health. *The Lancet, 370*(9591), 918-919.

Bass, J. K., Ryder, R. W., Lammers, M., Mukaba, T. N., & Bolton, P. A. (2008). Post-partum depression in Kinshasa, democratic republic of Congo: Validation of a concept using a mixed-methods cross-cultural approach. *Tropical Medicine & International Health, 13*(12), 1534-1542.

Bolton, P., Lee, C., Haroz, E. E., Murray, L., Dorsey, S., Robinson, C., . . . Bass, J. (2014). A transdiagnostic community-based mental health treatment for comorbid disorders: Development and outcomes of a randomized controlled trial among Burmese refugees in Thailand. *PLoS Medicine, 11*(11), e1001757.

Brownhill, S., Wilhelm, K., Barclay, L., & Schmied, V. (2005). 'Big build': Hidden depression in men. *Australian and New Zealand Journal of Psychiatry, 39*(10), 921-931.

Cochran, S. V., & Rabinowitz, F. E. (2003). Gender-sensitive recommendations for assessment and treatment of depression in men. *Professional Psychology: Research and Practice, 34*(2), 132-140.

De Jong, J. T. V. M., & Van Ommeren, M. (2002). Toward a culture-informed epidemiology: Combining qualitative and quantitative research in transcultural contexts. *Transcultural Psychiatry, 39*(4), 422-433.

Draguns, J. G., & Tanaka-Matsumi, J. (2003). Assessment of psychopathology across and within cultures: Issues and findings. *Behaviour Research and Therapy, 41*(7), 755-776.

Hambleton, R. K., Waminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (1st ed.). California: Sage Publications Inc.

Haroz, E. E., Bass, J. K., Lee, C., Murray, L. K., Robinson, C., & Bolton, P. (2014). Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the Thailand Burma border. *BMC Psychology, 2*(1), 31-40.

Hinton, D. E., Reis, R., & Jong, J. (2015). The "Thinking a lot" idiom of distress and PTSD: An examination of their relationship among traumatized Cambodian refugees using the "Thinking a lot" questionnaire. *Medical Anthropology Quarterly.*

Hollifield, M. (2002). Accurate measure in cultural psychiatry: Will we pay the costs? *Transcultural Psychiatry, 39*, 419–421.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual Review of Public Health, 34*(1), 119-138.

Kleinman, A. M. (1977). Depression, somatization and the "new cross-cultural psychiatry". *Social Science & Medicine (1967), 11*(1), 3-9.

Kohrt, B. A., Jordans, M. J. D., Tol, W. A., Luitel, N. P., Maharjan, S. M., & Upadhaya, N. (2011). Validation of cross-cultural child mental health and psychosocial research instruments: Adapting the depression self-rating scale and child PTSD symptom scale in nepal. *BMC Psychiatry, 11-27.*

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*(9), 1-7.

Nock, M. K., Borges, G., Bromet, E. J., Alonso, J., Angermeyer, M., Beautrais, A., . . . Williams, D. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry, 192*(2), 98-105.

Rasmussen, A., Keatley, E., & Joscelyne, A. (2014). Posttraumatic stress in emergency settings outside north america and europe: A review of the emic literature. *Social Science & Medicine, 109*, 44-54.

Simon, G., Goldberg, D., Von Korff, M., & Üstün, T. (2002). Understanding cross-national differences in depression prevalence. *Psychological Medicine, 32*(04), 585-594.

Simon, G., Gater, R., Kisely, S., & Piccinelli, M. (1996). Somatic symptoms of distress: An international primary care study. *Psychosomatic Medicine, 58*(5), 481-488.

Steel, Z., Chey, T., Silove, D., Marnane, C., Bryant, R. A. & van Ommeren, M. (2009). Association of torture and other potentially traumatic events with mental health outcomes among populations exposed to mass conflict and displacement: A systematic review and meta-analysis. *JAMA, 302*(5), 537-549.

Uebelacker, L., Strong, D., Weinstock, L., & Miller, I. (2009). Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. *Psychological Medicine, 39*(04), 591-601.

World Health Organization. (2011). *Mental health atlas, 2011.* Geneva: World Health Organization.

This page intended to be left blank

Appendix A. PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	

Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and	

		measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

Appendix B. All articles included in systematic review of qualitative studies related to depression

Table 1.
Studies reviewed (*N* = 106)

Author	Number of study populations specified (if more than 1)	Ethnicity/nationality	Sample type
Abas et al. (1994)		Zimbabwean	Health workers
Abbo et al. (2008)		Ugandan	Key informants
Abbot & Klein (1979)		Kenyan	Community
Kadir & Bifulco (2010)		Malaysian	Community
Abrams & Curran (2011)		American	Community
Amankwaa (2003)		American	Clinical
Andjajani-Sutahjo et al. (2007)		Indonesian	Clinical
Avotri & Walters, (1999)		Ghanian	Community
Bass et al. (2008)		Congolese	Key informant
Beiser et al. (1994)	3	Chinese, Vietnamese, Laotian	Refugee
Beiser et al. (1976)		Senegalese	Community
Berstein et al. (2008)		Korean	Community
Bolton (2001)		Rwandan	Key informant, community
Bolton et al. (2012)		Haitian	Key informant
Bolton (2013)		Kurdish	Key informant
Borra (2011)		Turkish	Clinical, community
Brown et al. (2012)		Aboriginee	Community, traditional healers
Brownhill et al. (2002)		Australian	Community
Bryant-Bedell & Waite (2010)		American	Clinical
Burr & Chapman (2004)		South Asian	Community
Cabassa et al. (2008)		American	Clinical
Cardozo et al. (2004)		Burmese	Refugee
Chan et al. (2002)		Chinese	Clinical
Chao (2011)		American	Community
Chen et al. (2002)		American	Community
Cortes (2003)		Puerto Rican	
Csordas et al. (2008)		American Indian	Community
Danielsson et al. (2011)		Swedish	Clinical
Dejman (2011)		Iranian	Clinical
Edhborg et al. (2005)		Swedish	Community
Etowa et al. (2007)		Canadian	Community
Familiar et al. (2013)		Burundian	Community
Farias (1991)		American	Refugee
Fenton & Sadiq-Sangster (1996)		South Asian	Community
Fox (2003)		Gambian	Traditional healers
Gao et al. (2010)		Chinese	Clinical
Ghubash & Eapen (2009)		UAE	Community, health workers
Halbreich et al. (2007)	9	Indian, Brazilian,	Community, mental

		Peruvian, Chilean, Venezuelan, Moroccan, Tunisian Serbian Hungarian	health professionals
Hanley (2007)		Bangladeshi	Community
Hanlon (2010)		Ethiopian	Community, health workers, traditional healers
Hung et al. (2006)		Taiwanese	Clinical
Jackson et al. (2008)		Canadian	HIV positive community
Jadhav et al. (2001)		British	Clinical
James et al. (2005)		Azores	Community, immigrant
Jayawickreme et al. (2009)		Sri Lankan	Community
Kaaya et al. (2010)		Tanzanian	Community, traditional healers
Kadem et al. (2001)		British	Clinical
Kaiser et al. (2013)		Haitian	Community
Karasz 2009		American	Clinical
Karasz (2005)		Indian	Community, immigrant
Kay (1989)		Mexican	Community
Kemp (2003)	2	St. Helenian, British	Key informants
Kendrick et al. (2007)		American	Community
Keys et al. (2012)		Haitian	Key informants
Koo (2012)		Chinese	Caregivers
Lackey (2008)		American	Community, immigrant
Lazear et al. (2008)		American	Community
Lee et al. (2011)		Burmese	Refugee
Lee et al. (2000)		Taiwanese	Community
Lee et al. (2007)		Chinese	Clinical
Liang & George (2012)		Indian	Community
Lim et al. (2013)		Burmese	Health workers
Mallinson & Popay (2007)		English	Community
Martinez et al. (2011)	4	Colombian, Cuban, Puerto Rican, Mexican	Community, immigrant
Meffert & Marmar (2009)		Sudanese	Refugee
Miller et al. (2006)		Afghani	Community
Mosotho (2008)		South African	Clinical
Muhwesi (2008)		Ugandan	Caregivers
Mumford et al. (2005)		Pakistani	Community, clinical
Murray et al. (2006)		Zambian	Key informants
Naeem et al. (2012)		Pakistani	Clinical
Nakimuli Mpungu et al. (2012)		Ugandan	HIV positive community
Nazroo (2002)	4	Bangladeshi, Pakistani, Afro-Caribbean, Irish	Community
Nichter (1981)		Indian	Community

Nieuwsma (2011)	2	American, Indian	Community
Núñez (2009)		Peruvian	Community, immigrant
Okello & Neema, (2007)		Ugandan	Clinical
Okello et al. (2012)		Ugandan	HIV positive community
Parker et al. (2001)		Chinese	Mental health professionals
Patel et al. (1995)		Zimbabwean	Community
Patel et al. (1997)		Zimbabwean	Community
Pereira et al. (2007)		Indian	Community
Phan et al. (2004)		Vietnamese	Refugee, immigrant
Pincay & Guarnaccia (2007)	4	Puerto Rican, Dominican, Mexican, Cuban	Community
Poleshuck et al. (2013)		American	Community
Poudyal et al. (2009)		Indonesian	Key informants
Hinton (2010)		papau new guinea	Key informants
Ranguram et al. (2001)		Indian	Clinical
Rao et al. (2012)		Indian	Clinical
Rasmussen et al. (2014)		Haitian	Clinical
Rees & Silove (2011)		West Papaun	Refugee
Rodrigues et al. (2003})		Indian	Community
Selim (2010)		Bangladeshi	Clinical, community, healthcare workers
Sellers et al. (2006)	3	Ghanaian, Cameroonian, Nigerian	Community
Shankar et al. (2006)		Indian	Traditional healers
Shin (2010)		Korean	Clinical
Sin et al. (2011)		Korean	Community
Sulaiman et al. (2001)		UAE	Community
Templeton et al. (2003)	3	Bangladeshi, Indian, Portuguese	Community
Tilbury (2007)	4	Somali, Ethiopian, Eritrean, Sudanese	Community
Ventevogel et al. (2013)	3	Burundian, South Sudanese, Congolese	Community
Waite & Killian, (2009)		American	Community
Walters et al. (1999)	2	English, Ghanaian	Community
White (2004)		Cambodian	Community, birth attendants
Wilk & Bolton, (2002)		Ugandan	Key informants
Youngmann et al. (1999)		Ethiopian	Service providers

References

Abas, M., Broadhead, J., Mbape, P., & Khumalo-Sakatukwa, G. (1994). Defeating depression in the developing world: A zimbabwean model. *The British Journal of Psychiatry, 164*(3), 293-296.

Abbo, C., Okello, E., Ekblad, S., Waako, P., & Musisi, S. (2008). Lay concepts of psychosis in busoga, eastern uganda: A pilot study. *J World Cultural Psychiatry Research Review, 3*(3), 132-145.

Abbott, S., & Klein, R. (1979). Depression and anxiety among rural Kikuyu in Kenya. *Ethos, 7*(2), 161-188.

Abrams, L. S., & Curran, L. (2011). Maternal identity negotiations among low-income women with symptoms of postpartum depression. *Qualitative Health Research, 3,* 373-385.

Amankwaa, L. C. (2003). Postpartum depression, culture and African-American women. *Journal of Cultural Diversity, 10*(1), 23-29.

Andajani-Sutjahjo, S., Manderson, L., & Astbury, J. (2007). Complex emotions, complex problems: Understanding the experiences of perinatal depression among new mothers in urban indonesia. *Culture, Medicine and Psychiatry, 31*(1), 101-122.

Avotri, J. Y., & Walters, V. (1999). ``You just look at our work and see if you have any freedom on earth": Ghanaian women's accounts of their work and their health. *Social Science & Medicine, 48*(9), 1123-1133.

Bass, J. K., Ryder, R. W., Lammers, M. C., Mukaba, T. N., & Bolton, P. (2008). Post-partum depression in Kinshasa, Democratic Republic of Congo: Validation of a concept using a

mixed-methods cross-cultural approach. *Tropical Medicine & International Health, 13*(12), 1534-1542.

Beiser, M., Benfari, R. C., Collomb, H., & Ravel, J. L. (1976). Measuring psychoneurotic behavior in cross-cultural surveys. *The Journal of Nervous and Mental Disease, 163*(1), 10-23.

Beiser, M., Cargo, M., & Woodbury, M. (1994). A comparison of psychiatric-disorder in different cultures - depressive typologies in Southeast-Asian refugees and resident Canadians. *International journal of methods in psychiatric research, 4*(3), 157-172.

Bernstein, K. S., Lee, J., Park, S., & Jyoung, J. (2008). Symptom manifestations and expressions among korean immigrant women suffering with depression. *Journal of Advanced Nursing, 61*(4), 393-402.

Bolton, P. (2001). Local perceptions of the mental health effects of the Rwandan genocide. *The Journal of Nervous and Mental Disease, 189*(4), 243-248.

Bolton, P., Michalopoulos, L., Ahmed, A. M., Murray, L. K., & Bass, J. (2013). The mental health and psychosocial problems of survivors of torture and genocide in Kurdistan, northern Iraq: A brief qualitative study. *Torture, Quarterly Journal on Rehabilitation of Torture Victims and Prevention of Torture, 23*(1), 1-14.

Bolton, P., Surkan, P. J., Gray, A. E., & Desmousseaux, M. (2012). The mental health and psychosocial effects of organized violence: A qualitative study in northern Haiti. *Transcultural Psychiatry, 49*(3-4), 590-612.

Borra, R. (2011). Depressive disorder among Turkish women in the Netherlands: A qualitative study of idioms of distress. *Transcultural Psychiatry, 48*(5), 660-674.

Brown, A., Scales, U., Beever, W., Rickards, B., Rowley, K., & O'Dea, K. (2012). Exploring the expression of depression and distress in aboriginal men in central Australia: A qualitative study. *BMC Psychiatry, 12*(1), 97-109.

Brownhill, S., Wilhelm, K., Barclay, L. & Parker, G. (2002). Detecting depression in men: A matter of guesswork. *International Journal of Mens Health, 1*(3), 259-271.

Bryant-Bedell, K., & Waite, R. (2010). Understanding major depressive disorder among middle-aged African American men. *Journal of Advanced Nursing, 66*(9), 2050-2060.

Burr, J., & Chapman, T. (2004). Contextualising experiences of depression in women from south Asian communities: A discursive approach. *Sociology of Health & Illness, 26*(4), 433-452.

Cabassa, L. J., Hansen, M. C., Palinkas, L. A., & Ell, K. (2008). Azucar y nervios: Explanatory models and treatment experiences of Hispanics with diabetes and depression. *Social Science & Medicine (1982), 66*(12), 2413-2424.

Chan, S. W., Levy, V., Chung, T. K., & Lee, D. (2002). A qualitative study of the experiences of a group of Hong Kong Chinese women diagnosed with postnatal depression. *Journal of Advanced Nursing, 39*(6), 571-579.

Chao, R. C., & Green, K. E. (2011). Multiculturally sensitive mental health scale (MSMHS): Development, factor analysis, reliability, and validity. *Psychological Assessment, 23*(4), 876-887.

Chen, J. P., Chen, H., & Chung, H. (2002). Case-based reviews: Depressive disorders in Asian American adults. *Western Journal of Medicine, 176*(4), 239-244.

Cortés, D. E. (2003). Idioms of distress, acculturation, and depression: The Puerto Rican experience. In K. M. Chun, P. Balls Organista & G. Marín (Eds.), (pp. 207-222). Washington, DC US: American Psychological Association.

Csordas, T. J., Storck, M. J., & Strauss, M. (2008). Diagnosis and distress in Navajo healing. *The Journal of Nervous and Mental Disease, 196*(8), 585-596.

Danielsson, U. E., Bengs, C., Samuelsson, E., & Johansson, E. E. (2011). "My greatest dream is to be normal": The impact of gender on the depression narratives of young Swedish men and women. *Qualitative Health Research, 21*(5), 612-624.

Dejman, M., Forouzan, A. S., Assari, S., Malekafzali, H., Nohesara, S., Khatibzadeh, N., . . . Ekblad, S. (2011). An explanatory model of depression among female patients in Fars, Kurds, Turks ethnic groups of Iran. *Iranian Journal of Public Health, 40*(3), 79-88.

Edhborg, M., Friberg, M., Lundh, W., & Widstrom, A. M. (2005). "Struggling with life": Narratives from women with signs of postpartum depression. *Scandinavian Journal of Public Health, 33*(4), 261-267.

Etowa, J., Keddy, B., Egbeyemi, J., & Eghan, F. (2007). Depression: The 'invisible grey fog' influencing the midlife health of African Canadian women. *International Journal of Mental Health Nursing, 16*(3), 203-213.

Familiar, I., Sharma, S., Ndayisaba, H., Munyentwari, N., Sibomana, S., & Bass, J. K. (2013). Community perceptions of mental distress in a post-conflict setting: A qualitative study in Burundi. *Global Public Health, 8*(8), 943-957.

Farias, P. J. (1991). Emotional distress and its socio-political correlates in Salvadoran refugees: Analysis of a clinical sample. *Culture, Medicine and Psychiatry, 15*(2), 167-192.

Fenton, S., & Sadiq-Sangster, A. (1996). Culture, relativism and the expression of mental distress: South Asian women in Britain. *Sociology of Health & Illness, 18*(1), 66-85.

Fox, S. H. (2003). The Mandinka nosological system in the context of post-trauma syndromes. *Transcultural Psychiatry, 40*(4), 488-506.

Gao, L. L., Chan, S. W., You, L., & Li, X. (2010). Experiences of postpartum depression among first-time mothers in mainland China. *Journal of Advanced Nursing, 66*(2), 303-312.

Ghubash, R., & Eapen, V. (2009). Postpartum mental illness: Perspectives from an Arabian gulf population. *Psychological Reports, 105*(1), 127-136.

Halbreich, U., Alarcon, R. D., Calil, H., Douki, S., Gaszner, P., Jadresic, E., . . . Trivedi, J. K. (2007). Culturally-sensitive complaints of depressions and anxieties in women. *Journal of Affective Disorders, 102*(1-3), 159-176.

Hanley, J. (2007). The emotional wellbeing of Bangladeshi mothers during the postnatal period. *Community Practitioner: The Journal of the Community Practitioners' & Health Visitors' Association, 80*(5), 34-37.

Hanlon, C., Whitley, R., Wondimagegn, D., Alem, A., & Prince, M. (2010). Between life and death: Exploring the sociocultural context of antenatal mental distress in rural Ethiopia. *Archives of Women's Mental Health, 13*(5), 385-393.

Hinton, R., & Earnest, J. (2010). 'I worry so much I think it will kill me': Psychosocial health and the links to the conditions of women's lives in Papua New Guinea. *Health Sociology Review, 19*(1), 5-19.

Hung, C. I., Weng, L. J., Su, Y. J., & Liu, C. Y. (2006). Preliminary study of a scale measuring depression and somatic symptoms. *Psychological Reports, 99*(2), 379-389.

Jackson, R., Prentice, T., Collins, E., & Mill, J. (2008). *Depression among aboriginal people living with HIV/AIDS.* Online. Available: http://caan.ca/wp-content/uploads/2012/05/Depression-Among-APHAs-EN.pdf

Jadhav, S., Weiss, M. G., & Littlewood, R. (2001). Cultural experience of depression among white Britons in London. *Anthropology & Medicine, 8*(1), 47-70.

James, S., Navara, G. S., Clarke, J. N., & Lomotey, J. (2005). An inquiry into the "agonies"(agonias) of Portuguese immigrants from the Azores. *Hispanic Journal of Behavioral Sciences, 27*(4), 547-564.

Jayawickreme, N., Jayawickreme, E., Goonasekera, M. A., & Foa, E. B. (2009). Distress, wellbeing and war: Qualitative analyses of civilian interviews from north eastern Sri Lanka. *Intervention, 7*(3), 204-222.

Kaaya, S. F., Mbwambo, J. K., Smith Fawzi, M. C., Van, D. B., Schaalma, H., & Leshabari, M. T. (2010). Understanding women's experiences of distress during pregnancy in Dar es Salaam, Tanzania. *Tanzania Journal of Health Research, 12*(1), 36-46.

Kadam, U. T., Croft, P., McLeod, J., & Hutchinson, M. (2001). A qualitative study of patients' views on anxiety and depression. *The British Journal of General Practice, 51*(466), 375-380.

Kadir, N. B. A., & Bifulco, A. (2010). Malaysian moslem mothers' experience of depression and service use. *Culture, Medicine and Psychiatry, 34*(3), 443-467.

Kaiser, B. N., Kohrt, B. A., Keys, H. M., Khoury, N. M., & Brewster, A. R. (2013). Strategies for assessing mental health in Haiti: Local instrument development and transcultural translation. *Transcultural Psychiatry, 50*(4), 532-558.

Karasz, A. (2005). Cultural differences in conceptual models of depression. *Social Science & Medicine,* 60(7), 1625-1635.

Kemp, M. (2003). Hearts and minds: Agency and discourse on distress. *Anthropology & Medicine, 10*(2), 187-205.

Kendrick, L., Anderson, N. L. R. & Moore, B. (2007). Perceptions of depression among young African American men. *Family Community Health, 30*(1), 63-73.

Keys, H. M., Kaiser, B. N., Kohrt, B. A., Khoury, N. M., & Brewster, A. R. T. (2012). Idioms of distress, ethnopsychology, and the clinical encounter in Haiti's central plateau. *Social Science & Medicine, 75*(3), 555-564.

Koo, K. (2012). Carers' representations of affective mental disorders in British Chinese communities. *Sociology of Health and Illness, 34*(8), 1140-1155.

Lackey, G. F. (2008). "Feeling blue" in Spanish: A qualitative inquiry of depression among Mexican immigrants. *Social Science and Medicine, 67*(2), 228-237.

Lazear, K. J., Pires, S. A., Isaacs, M. R., Chaulk, P., & Huang, L. (2008). Depression among low-income women of color: Qualitative findings from cross-cultural focus groups. *Journal of Immigrant and Minority Health, 10*(2), 127-133.

Lee, C., Robinson, C., & Bolton, P. (2011). Qualitative assessment of displaced persons in Mae Sot, Thailand affected by torture and related violence in Burma. *A report*

to the Victims of Torture Fund, United States Agency for International Development (USAID).

Lee, D. T., Kleinman, J., & Kleinman, A. (2007). Rethinking depression: An ethnographic study of the experiences of depression among Chinese. *Harvard Review of Psychiatry, 15*(1), 1-8.

Lee, Y., Yang, M. J., Lai, T. J., Chiu, N. M., & Chau, T. T. (2000). Development of the Taiwanese depression questionnaire. *Chang Gung Medical Journal, 23*(11), 688-694.

Liang, T. K., & George, T. S. (2012). Men's experiences of depression and the family's role in gender socialization: A phenomenological study from urban south India. *Journal of Comparative Family Studies, 43*(1), 93-131.

Lim, A. G., Stock, L., Shwe Oo, E. K., & Jutte, D. P. (2013). Trauma and mental health of medics in eastern Myanmar's conflict zones: A cross-sectional and mixed methods investigation. *Conflict and Health, 7*(1), 15-28.

Lopes Cardozo, B., Talley, L., Burton, A., & Crawford, C. (2004). Karenni refugees living in Thai–Burmese border camps: Traumatic experiences, mental health outcomes, and social functioning. *Social Science & Medicine, 58*(12), 2637-2644.

Mallinson, S., & Popay, J. (2007). Describing depression: Ethnicity and the use of somatic imagery in accounts of mental distress. *Sociology of Health & Illness, 29*(6), 857-871.

Martinez Tyson, D. D., Castañeda, H., Porter, M., Quiroz, M., & Carrion, I. (2011). More similar than different? Exploring cultural models of depression among Latino immigrants in Florida. *Depression Research and Treatment,* 1-11.

Meffert, S. M., & Marmar, C. R. (2009). Darfur refugees in Cairo mental health and interpersonal conflict in the aftermath of genocide. *Journal of Interpersonal Violence, 24*(11), 1835-1848.

Miller, K. E., Omidian, P., Quraishy, A. S., Quraishy, N., Nasiry, M. N., Nasiry, S., . . . Yaqubi, A. A. (2006). The afghan symptom checklist: A culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry, 76*(4), 423-433.

Mosotho, N., Louw, D., Calitz, F., & Esterhuyse, K. (2008). Depression among Sesotho speakers in Mangaung, South Africa. *African Journal of Psychiatry, 11*(1), 35-43.

Mumford, D. B., Ayub, M., Karim, R., Izhar, N., Asif, A., & Bavington, J. T. (2005). Development and validation of a questionnaire for anxiety and depression in Pakistan. *Journal of Affective Disorders, 88*(2), 175-182.

Murray, L., Bass, J., & Bolton, P. (2006). Qualitative study to identify indicators of psychological problems and functional impairment among residents of Sange district, south Kivu, eastern DRC. *A report to the Victims of Torture Fund, United States Agency for International Development (USAID).*

Naeem, F., Ayub, M., Kingdon, D., & Gobbi, M. (2012). Views of depressed patients in Pakistan concerning their illness, its causes, and treatments. *Qualitative Health Research, 22*(8), 1083-1093.

Nakimuli-Mpungu, E., Mojtabai, R., Alexandre, P. K., Katabira, E., Musisi, S., Nachega, J. B., & Bass, J. K. (2012). Cross-cultural adaptation and validation of the self-reporting questionnaire among HIV individuals in a rural ART program in southern Uganda. *HIV AIDS (Auckl), 4,* 51-60.

Nazroo, J., & O'Connor, W. (2002). *Idioms of mental distress.* London: National Centre for Social Research.

Nichter, M. (1981). Idioms of distress: Alternatives in the expression of psychosocial distress: A case study from south India. *Culture, Medicine and Psychiatry, 5*(4), 379-408.

Nieuwsma, J. A. (2011). Indigenous perspectives on depression in rural regions of India and the United States. *Transcultural Psychiatry, 48*(5), 539-568.

Núñez, L. (2009). Is it possible to eradicate poverty without attending to mental health? Listening to migrant workers in Chile through their idioms of distress. *Journal of Health Management, 11*(2), 337-354.

Okello, E. S., & Neema, S. (2007). Explanatory models and help-seeking behavior: Pathways to psychiatric care among patients admitted for depression in Mulago hospital, Kampala, Uganda. *Qualitative Health Research, 17*(1), 14-25.

Okello, E. S., Ngo, V. K., Ryan, G., Musisi, S., Akena, D., Nakasujja, N., & Wagner, G. (2012). Qualitative study of the influence of antidepressants on the psychological health of patients on antiretroviral therapy in Uganda. *African Journal of AIDS research, 11*(1), 27-44.

Parker, G., Gladstone, G., & Chee, K. T. (2001). Depression in the planet's largest ethnic group: The Chinese. *The American Journal of Psychiatry, 158*(6), 857-864.

Patel, V., Simunyu, E., Gwanzura, F., Lewis, G., & Mann, A. (1997). The Shona symptom questionnaire: The development of an indigenous measure of common mental disorders in harare. *Acta Psychiatrica Scandinavica, 95*(6), 469-475.

Patel, V. (1995). Explanatory models of mental illness in Sub-Saharan Africa. *Social Science & Medicine, 40*(9), 1291-1298.

Pereira, B., Andrew, G., Pai, R., Pelto, P., & Patel, V. (2007). The explanatory models of depression in low income countries: Listening to women in India. *Journal of Affective Disorders, 102*(1), 209-218.

Phan, T., Steel, Z., & Silove, D. (2004). An ethnographically derived measure of anxiety, depression and somatization: The Phan Vietnamese psychiatric scale. *Transcultural Psychiatry, 41*(2), 200-232.

Pincay, I. E. M., & Guarnaccia, P. J. (2007). 'It's like going through an earthquake': Anthropological perspectives on depression among latino immigrants. *Journal of Immigrant and Minority Health, 9*(1), 17-28.

Poleshuck, E. L., Cerrito, B., Leshoure, N., Finocan-Kaag, G., & Kearney, M. H. (2013). Underserved women in a women's health clinic describe their experiences of depressive symptoms and why they have low uptake of psychotherapy. *Community Mental Health Journal, 49*(1), 50-60.

Poudyal, B., Bass, J., Subyantoro, T., Jonathan, A., Erni, T., & Bolton, P. (2009). Assessment of the psychosocial and mental health needs, dysfunction and coping mechanisms of violence affected populations in Bireuen, Aceh. *Torture: Quarterly Journal on Rehabilitation of Torture Victims and Prevention of Torture, 19*(3), 218-226.

Raguram, R., Weiss, M. G., Keval, H., & Channabasavanna, S. M. (2001). Cultural dimensions of clinical depression in Bangalore, India. *Anthropology & Medicine, 8*(1), 31-46.

Rao, D., Horton, R., & Raguram, R. (2012). Gender inequality and structural violence among depressed women in south India. *Social Psychiatry and Psychiatric Epidemiology, 47*(12), 1967-1975.

Rasmussen, A., Eustache, E., Raviola, G., Kaiser, B., Grelotti, D. J., & Belkin, G. S. (2014). Development and validation of a Haitian creole screening instrument for depression. *Transcultural Psychiatry, 52*(1), 33-57.

Rees, S., & Silove, D. (2011). Sakit Hati: A state of chronic mental distress related to resentment and anger amongst West Papuan refugees exposed to persecution. *Social Science & Medicine, 73*(1), 103-110.

Rodrigues, M., Patel, V., Jaswal, S., & de Souza, N. (2003). Listening to mothers: Qualitative studies on motherhood and depression from Goa, India. *Social Science & Medicine,* 57(10), 1797-1806.

Selim, N. (2010). Cultural dimensions of depression in Bangladesh: A qualitative study in two villages of Matlab. *Journal of Health, Population, and Nutrition, 28*(1), 95-106.

Sellers, S. L., Ward, E. C., & Pate, D. (2006). Dimensions of depression: A qualitative study of wellbeing among black African immigrant women. *Qualitative Social Work: Research and Practice, 5*(1), 45-66.

Shankar, B. R., Saravanan, B., & Jacob, K. S. (2006). Explanatory models of common mental disorders among traditional healers and their patients in rural south India. *The International Journal of Social Psychiatry, 52*(3), 221-233.

Shin, J. K. (2010). Understanding the experience and manifestation of depression among korean immigrants in new york city. *Journal of Transcultural Nursing, 21*(1), 73-80.

Sin, M. K., Jordan, P., & Park, J. (2011). Perceptions of depression in Korean American immigrants. *Issues in Mental Health Nursing, 32*(3), 177-183.

Sulaiman, S. O. Y., Bhugra, D., & De Silva, P. (2001). The development of a culturally sensitive symptom checklist for depression in Dubai. *Transcult.Psychiatry, 38*(2), 219-229.

Templeton, L., Velleman, R., Persaud, A., & Milner, P. (2003). The experiences of postnatal depression in women from black and minority ethnic communities in Wiltshire, UK. *Ethnicity and Health, 8*(3), 207-221.

Tilbury, F. (2007). "I feel I am a bird without wings": Discourses of sadness and loss among east Africans in western Australia. *Identities, 14*(4), 433-458.

Ventevogel, P., Jordans, M., Reis, R., & de Jong, J. (2013). Madness or sadness? Local concepts of mental illness in four conflict-affected African communities. *Conflict and Health, 7*(1), 3-19.

Waite, R., & Killian, P. (2009). Perspectives about depression: Explanatory models among African-American women. *Archives of Psychiatric Nursing, 23*(4), 323-333.

Walters, V., Avotri, J. Y., & Charles, N. (1999). "Your heart is never free": Women in wales and Ghana talking about distress. *Canadian Psychology/Psychologie Canadienne, 40*(2), 129-142.

White, P. (2004). Heat, balance, humors, and ghosts: Postpartum in Cambodia. *Health Care for Women International, 25*(2), 179-194.

Wilk, C. M., & Bolton, P. (2002). Local perceptions of the mental health effects of the Uganda acquired immunodeficiency syndrome epidemic. *The Journal of Nervous and Mental Disease, 190*(6), 394-397.

Youngmann, R., Minuchin-Itzigsohn, S., & Barasch, M. (1999). Manifestations of emotional distress among Ethiopian immigrants in Israel: Patient and clinician perspectives. *Transcultural Psychiatry, 36*(1), 45-63.

Appendix A. PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	85
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	86
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	87-89
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	90
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	n/a
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	91
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	90
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	90-91

Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	91
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	91
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	91
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	n/a
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	92
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	92

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	n/a
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	n/a
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	93-95
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	96-97
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	n/a

Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	n/a
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	97-103
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	n/a
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	n/a
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	103-106
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	106-107
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	103-106
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	v

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

Appendix B. Articles included in systematic review of qualitative studies related to depression

Table 1.
Studies reviewed (*N* = 106)

Author	Number of study populations specified (if more than 1)	Ethnicity/ nationality	Sample type
Abas et al. (1994)		Zimbabwean	Health workers
Abbo et al. (2008)		Ugandan	Key informants
Abbot & Klein (1979)		Kenyan	Community
Kadir & Bifulco (2010)		Malaysian	Community
Abrams & Curran (2011)		American	Community
Amankwaa (2003)		American	Clinical
Andjajani-Sutahjo et al. (2007)		Indonesian	Clinical
Avotri & Walters, (1999)		Ghanian	Community
Bass et al. (2008)		Congolese	Key informant
Beiser et al. (1994)	3	Chinese, Vietnamese, Laotian	Refugee
Beiser et al. (1976)		Senegalese	Community
Berstein et al. (2008)		Korean	Community
Bolton (2001)		Rwandan	Key informant, community
Bolton et al. (2012)		Haitian	Key informant
Bolton (2013)		Kurdish	Key informant
Borra (2011)		Turkish	Clinical, community
Brown et al. (2012)		Aboriginee	Community, traditional healers
Brownhill et al. (2002)		Australian	Community
Bryant-Bedell & Waite (2010)		American	Clinical
Burr & Chapman (2004)		South Asian	Community
Cabassa et al. (2008)		American	Clinical
Cardozo et al. (2004)		Burmese	Refugee
Chan et al. (2002)		Chinese	Clinical
Chao (2011)		American	Community
Chen et al. (2002)		American	Community
Cortes (2003)		Puerto Rican	
Csordas et al. (2008)		American Indian	Community
Danielsson et al. (2011)		Swedish	Clinical
Dejman (2011)		Iranian	Clinical
Edhborg et al. (2005)		Swedish	Community
Etowa et al. (2007)		Canadian	Community
Familiar et al. (2013)		Burundian	Community
Farias (1991)		American	Refugee

Fenton & Sadiq-Sangster (1996)		South Asian	Community
Fox (2003)		Gambian	Traditional healers
Gao et al. (2010)		Chinese	Clinical
Ghubash & Eapen (2009)		UAE	Community, health workers
Halbreich et al. (2007)	9	Indian, Brazilian, Peruvian, Chilean, Venezuelan, Moroccan, Tunisian Serbian Hungarian	Community, mental health professionals
Hanley (2007)		Bangladeshi	Community
Hanlon (2010)		Ethiopian	Community, health workers, traditional healers
Hung et al. (2006)		Taiwanese	Clinical
Jackson et al. (2008)		Canadian	HIV positive community
Jadhav et al. (2001)		British	Clinical
James et al. (2005)		Azores	Community, immigrant
Jayawickreme et al. (2009)		Sri Lankan	Community
Kaaya et al. (2010)		Tanzanian	Community, traditional healers
Kadem et al. (2001)		British	Clinical
Kaiser et al. (2013)		Haitian	Community
Karasz 2009		American	Clinical
Karasz (2005)		Indian	Community, immigrant
Kay (1989)		Mexican	Community
Kemp (2003)	2	St. Helenian, British	Key informants
Kendrick et al. (2007)		American	Community
Keys et al. (2012)		Haitian	Key informants
Koo (2012)		Chinese	Caregivers
Lackey (2008)		American	Community, immigrant
Lazear et al. (2008)		American	Community
Lee et al. (2011)		Burmese	Refugee
Lee et al. (2000)		Taiwanese	Community
Lee et al. (2007)		Chinese	Clinical
Liang & George (2012)		Indian	Community
Lim et al. (2013)		Burmese	Health workers
Mallinson & Popay (2007)		English	Community
Martinez et al. (2011)	4	Colombian,	Community,

		Cuban, Puerto Rican, Mexican	immigrant
Meffert & Marmar (2009)		Sudanese	Refugee
Miller et al. (2006)		Afghani	Community
Mosotho (2008)		South African	Clinical
Muhwesi (2008)		Ugandan	Caregivers
Mumford et al. (2005)		Pakistani	Community, clinical
Murray et al. (2006)		Zambian	Key informants
Naeem et al. (2012)		Pakistani	Clinical
Nakimuli Mpungu et al. (2012)		Ugandan	HIV positive community
Nazroo (2002)	4	Bangladeshi, Pakistani, Afro-Caribbean, Irish	Community
Nichter (1981)		Indian	Community
Nieuwsma (2011)	2	American, Indian	Community
Núñez (2009)		Peruvian	Community, immigrant
Okello & Neema, (2007)		Ugandan	Clinical
Okello et al. (2012)		Ugandan	HIV positive community
Parker et al. (2001)		Chinese	Mental health professionals
Patel et al. (1995)		Zimbabwean	Community
Patel et al. (1997)		Zimbabwean	Community
Pereira et al. (2007)		Indian	Community
Phan et al. (2004)		Vietnamese	Refugee, immigrant
Pincay & Guarnaccia (2007)	4	Puerto Rican, Dominican, Mexican, Cuban	Community
Poleshuck et al. (2013)		American	Community
Poudyal et al. (2009)		Indonesian	Key informants
Hinton (2010)		papau new guinea	Key informants
Ranguram et al. (2001)		Indian	Clinical
Rao et al. (2012)		Indian	Clinical
Rasmussen et al. (2014)		Haitian	Clinical
Rees & Silove (2011)		West Papaun	Refugee
Rodrigues et al. (2003})		Indian	Community
Selim (2010)		Bangladeshi	Clinical, community, healthcare workers
Sellers et al. (2006)	3	Ghanaian, Cameroonian, Nigerian	Community

Shankar et al. (2006)		Indian	Traditional healers
Shin (2010)		Korean	Clinical
Sin et al. (2011)		Korean	Community
Sulaiman et al. (2001)		UAE	Community
Templeton et al. (2003)	3	Bangladeshi, Indian, Portuguese	Community
Tilbury (2007)	4	Somali, Ethiopian, Eritrean, Sudanese	Community
Ventevogel et al. (2013)	3	Burundian, South Sudanese, Congolese	Community
Waite & Killian, (2009)		American	Community
Walters et al. (1999)	2	English, Ghanaian	Community
White (2004)		Cambodian	Community, birth attendants
Wilk & Bolton, (2002)		Ugandan	Key informants
Youngmann et al. (1999)		Ethiopian	Service providers

References

Abas, M., Broadhead, J., Mbape, P., & Khumalo-Sakatukwa, G. (1994). Defeating depression in the developing world: A zimbabwean model. *The British Journal of Psychiatry, 164*(3), 293-296.

Abbo, C., Okello, E., Ekblad, S., Waako, P., & Musisi, S. (2008). Lay concepts of psychosis in busoga, eastern uganda: A pilot study. *J World Cultural Psychiatry Research Review, 3*(3), 132-145.

Abbott, S., & Klein, R. (1979). Depression and anxiety among rural Kikuyu in Kenya. *Ethos, 7*(2), 161-188.

Abrams, L. S., & Curran, L. (2011). Maternal identity negotiations among low-income women with symptoms of postpartum depression. *Qualitative Health Research, 3,* 373-385.

Amankwaa, L. C. (2003). Postpartum depression, culture and African-American women. *Journal of Cultural Diversity, 10*(1), 23-29.

Andajani-Sutjahjo, S., Manderson, L., & Astbury, J. (2007). Complex emotions, complex problems: Understanding the experiences of perinatal depression among new mothers in urban indonesia. *Culture, Medicine and Psychiatry, 31*(1), 101-122.

Avotri, J. Y., & Walters, V. (1999). ``You just look at our work and see if you have any freedom on earth": Ghanaian women's accounts of their work and their health. *Social Science & Medicine, 48*(9), 1123-1133.

Bass, J. K., Ryder, R. W., Lammers, M. C., Mukaba, T. N., & Bolton, P. (2008). Post-partum depression in Kinshasa, Democratic Republic of Congo: Validation of a concept using a

mixed-methods cross-cultural approach. *Tropical Medicine & International Health, 13*(12), 1534-1542.

Beiser, M., Benfari, R. C., Collomb, H., & Ravel, J. L. (1976). Measuring psychoneurotic behavior in cross-cultural surveys. *The Journal of Nervous and Mental Disease, 163*(1), 10-23.

Beiser, M., Cargo, M., & Woodbury, M. (1994). A comparison of psychiatric-disorder in different cultures - depressive typologies in Southeast-Asian refugees and resident Canadians. *International journal of methods in psychiatric research, 4*(3), 157-172.

Bernstein, K. S., Lee, J., Park, S., & Jyoung, J. (2008). Symptom manifestations and expressions among korean immigrant women suffering with depression. *Journal of Advanced Nursing, 61*(4), 393-402.

Bolton, P. (2001). Local perceptions of the mental health effects of the Rwandan genocide. *The Journal of Nervous and Mental Disease, 189*(4), 243-248.

Bolton, P., Michalopoulos, L., Ahmed, A. M., Murray, L. K., & Bass, J. (2013). The mental health and psychosocial problems of survivors of torture and genocide in Kurdistan, northern Iraq: A brief qualitative study. *Torture, Quarterly Journal on Rehabilitation of Torture Victims and Prevention of Torture, 23*(1), 1-14.

Bolton, P., Surkan, P. J., Gray, A. E., & Desmousseaux, M. (2012). The mental health and psychosocial effects of organized violence: A qualitative study in northern Haiti. *Transcultural Psychiatry, 49*(3-4), 590-612.

Borra, R. (2011). Depressive disorder among Turkish women in the Netherlands: A qualitative study of idioms of distress. *Transcultural Psychiatry, 48*(5), 660-674.

Brown, A., Scales, U., Beever, W., Rickards, B., Rowley, K., & O'Dea, K. (2012). Exploring the expression of depression and distress in aboriginal men in central Australia: A qualitative study. *BMC Psychiatry, 12*(1), 97-109.

Brownhill, S., Wilhelm, K., Barclay, L. & Parker, G. (2002). Detecting depression in men: A matter of guesswork. *International Journal of Mens Health, 1*(3), 259-271.

Bryant-Bedell, K., & Waite, R. (2010). Understanding major depressive disorder among middle-aged African American men. *Journal of Advanced Nursing, 66*(9), 2050-2060.

Burr, J., & Chapman, T. (2004). Contextualising experiences of depression in women from south Asian communities: A discursive approach. *Sociology of Health & Illness, 26*(4), 433-452.

Cabassa, L. J., Hansen, M. C., Palinkas, L. A., & Ell, K. (2008). Azucar y nervios: Explanatory models and treatment experiences of Hispanics with diabetes and depression. *Social Science & Medicine (1982), 66*(12), 2413-2424.

Chan, S. W., Levy, V., Chung, T. K., & Lee, D. (2002). A qualitative study of the experiences of a group of Hong Kong Chinese women diagnosed with postnatal depression. *Journal of Advanced Nursing, 39*(6), 571-579.

Chao, R. C., & Green, K. E. (2011). Multiculturally sensitive mental health scale (MSMHS): Development, factor analysis, reliability, and validity. *Psychological Assessment, 23*(4), 876-887.

Chen, J. P., Chen, H., & Chung, H. (2002). Case-based reviews: Depressive disorders in Asian American adults. *Western Journal of Medicine, 176*(4), 239-244.

Cortés, D. E. (2003). Idioms of distress, acculturation, and depression: The Puerto Rican experience. In K. M. Chun, P. Balls Organista & G. Marín (Eds.), (pp. 207-222). Washington, DC US: American Psychological Association.

Csordas, T. J., Storck, M. J., & Strauss, M. (2008). Diagnosis and distress in Navajo healing. *The Journal of Nervous and Mental Disease, 196*(8), 585-596.

Danielsson, U. E., Bengs, C., Samuelsson, E., & Johansson, E. E. (2011). "My greatest dream is to be normal": The impact of gender on the depression narratives of young Swedish men and women. *Qualitative Health Research, 21*(5), 612-624.

Dejman, M., Forouzan, A. S., Assari, S., Malekafzali, H., Nohesara, S., Khatibzadeh, N., . . . Ekblad, S. (2011). An explanatory model of depression among female patients in Fars, Kurds, Turks ethnic groups of Iran. *Iranian Journal of Public Health, 40*(3), 79-88.

Edhborg, M., Friberg, M., Lundh, W., & Widstrom, A. M. (2005). "Struggling with life": Narratives from women with signs of postpartum depression. *Scandinavian Journal of Public Health, 33*(4), 261-267.

Etowa, J., Keddy, B., Egbeyemi, J., & Eghan, F. (2007). Depression: The 'invisible grey fog' influencing the midlife health of African Canadian women. *International Journal of Mental Health Nursing, 16*(3), 203-213.

Familiar, I., Sharma, S., Ndayisaba, H., Munyentwari, N., Sibomana, S., & Bass, J. K. (2013). Community perceptions of mental distress in a post-conflict setting: A qualitative study in Burundi. *Global Public Health, 8*(8), 943-957.

Farias, P. J. (1991). Emotional distress and its socio-political correlates in Salvadoran refugees: Analysis of a clinical sample. *Culture, Medicine and Psychiatry, 15*(2), 167-192.

Fenton, S., & Sadiq-Sangster, A. (1996). Culture, relativism and the expression of mental distress: South Asian women in Britain. *Sociology of Health & Illness, 18*(1), 66-85.

Fox, S. H. (2003). The Mandinka nosological system in the context of post-trauma syndromes. *Transcultural Psychiatry, 40*(4), 488-506.

Gao, L. L., Chan, S. W., You, L., & Li, X. (2010). Experiences of postpartum depression among first-time mothers in mainland China. *Journal of Advanced Nursing, 66*(2), 303-312.

Ghubash, R., & Eapen, V. (2009). Postpartum mental illness: Perspectives from an Arabian gulf population. *Psychological Reports, 105*(1), 127-136.

Halbreich, U., Alarcon, R. D., Calil, H., Douki, S., Gaszner, P., Jadresic, E., . . . Trivedi, J. K. (2007). Culturally-sensitive complaints of depressions and anxieties in women. *Journal of Affective Disorders, 102*(1-3), 159-176.

Hanley, J. (2007). The emotional wellbeing of Bangladeshi mothers during the postnatal period. *Community Practitioner: The Journal of the Community Practitioners' & Health Visitors' Association, 80*(5), 34-37.

Hanlon, C., Whitley, R., Wondimagegn, D., Alem, A., & Prince, M. (2010). Between life and death: Exploring the sociocultural context of antenatal mental distress in rural Ethiopia. *Archives of Women's Mental Health, 13*(5), 385-393.

Hinton, R., & Earnest, J. (2010). 'I worry so much I think it will kill me': Psychosocial health and the links to the conditions of women's lives in Papua New Guinea. *Health Sociology Review, 19*(1), 5-19.

Hung, C. I., Weng, L. J., Su, Y. J., & Liu, C. Y. (2006). Preliminary study of a scale measuring depression and somatic symptoms. *Psychological Reports, 99*(2), 379-389.

Jackson, R., Prentice, T., Collins, E., & Mill, J. (2008). *Depression among aboriginal people living with HIV/AIDS.* Online. Available: http://caan.ca/wp-content/uploads/2012/05/Depression-Among-APHAs-EN.pdf

Jadhav, S., Weiss, M. G., & Littlewood, R. (2001). Cultural experience of depression among white Britons in London. *Anthropology & Medicine, 8*(1), 47-70.

James, S., Navara, G. S., Clarke, J. N., & Lomotey, J. (2005). An inquiry into the "agonies"(agonias) of Portuguese immigrants from the Azores. *Hispanic Journal of Behavioral Sciences, 27*(4), 547-564.

Jayawickreme, N., Jayawickreme, E., Goonasekera, M. A., & Foa, E. B. (2009). Distress, wellbeing and war: Qualitative analyses of civilian interviews from north eastern Sri Lanka. *Intervention, 7*(3), 204-222.

Kaaya, S. F., Mbwambo, J. K., Smith Fawzi, M. C., Van, D. B., Schaalma, H., & Leshabari, M. T. (2010). Understanding women's experiences of distress during pregnancy in Dar es Salaam, Tanzania. *Tanzania Journal of Health Research, 12*(1), 36-46.

Kadam, U. T., Croft, P., McLeod, J., & Hutchinson, M. (2001). A qualitative study of patients' views on anxiety and depression. *The British Journal of General Practice, 51*(466), 375-380.

Kadir, N. B. A., & Bifulco, A. (2010). Malaysian moslem mothers' experience of depression and service use. *Culture, Medicine and Psychiatry, 34*(3), 443-467.

Kaiser, B. N., Kohrt, B. A., Keys, H. M., Khoury, N. M., & Brewster, A. R. (2013). Strategies for assessing mental health in Haiti: Local instrument development and transcultural translation. *Transcultural Psychiatry, 50*(4), 532-558.

Karasz, A. (2005). Cultural differences in conceptual models of depression. *Social Science & Medicine,* 60(7), 1625-1635.

Kemp, M. (2003). Hearts and minds: Agency and discourse on distress. *Anthropology & Medicine, 10*(2), 187-205.

Kendrick, L., Anderson, N. L. R. & Moore, B. (2007). Perceptions of depression among young African American men. *Family Community Health, 30*(1), 63-73.

Keys, H. M., Kaiser, B. N., Kohrt, B. A., Khoury, N. M., & Brewster, A. R. T. (2012). Idioms of distress, ethnopsychology, and the clinical encounter in Haiti's central plateau. *Social Science & Medicine, 75*(3), 555-564.

Koo, K. (2012). Carers' representations of affective mental disorders in British Chinese communities. *Sociology of Health and Illness, 34*(8), 1140-1155.

Lackey, G. F. (2008). "Feeling blue" in Spanish: A qualitative inquiry of depression among Mexican immigrants. *Social Science and Medicine, 67*(2), 228-237.

Lazear, K. J., Pires, S. A., Isaacs, M. R., Chaulk, P., & Huang, L. (2008). Depression among low-income women of color: Qualitative findings from cross-cultural focus groups. *Journal of Immigrant and Minority Health, 10*(2), 127-133.

Lee, C., Robinson, C., & Bolton, P. (2011). Qualitative assessment of displaced persons in Mae Sot, Thailand affected by torture and related violence in Burma. *A report to the Victims of Torture Fund, United States Agency for International Development (USAID).*

Lee, D. T., Kleinman, J., & Kleinman, A. (2007). Rethinking depression: An ethnographic study of the experiences of depression among Chinese. *Harvard Review of Psychiatry, 15*(1), 1-8.

Lee, Y., Yang, M. J., Lai, T. J., Chiu, N. M., & Chau, T. T. (2000). Development of the Taiwanese depression questionnaire. *Chang Gung Medical Journal, 23*(11), 688-694.

Liang, T. K., & George, T. S. (2012). Men's experiences of depression and the family's role in gender socialization: A phenomenological study from urban south India. *Journal of Comparative Family Studies, 43*(1), 93-131.

Lim, A. G., Stock, L., Shwe Oo, E. K., & Jutte, D. P. (2013). Trauma and mental health of medics in eastern Myanmar's conflict zones: A cross-sectional and mixed methods investigation. *Conflict and Health, 7*(1), 15-28.

Lopes Cardozo, B., Talley, L., Burton, A., & Crawford, C. (2004). Karenni refugees living in Thai–Burmese border camps: Traumatic experiences, mental health outcomes, and social functioning. *Social Science & Medicine, 58*(12), 2637-2644.

Mallinson, S., & Popay, J. (2007). Describing depression: Ethnicity and the use of somatic imagery in accounts of mental distress. *Sociology of Health & Illness, 29*(6), 857-871.

Martinez Tyson, D. D., Castañeda, H., Porter, M., Quiroz, M., & Carrion, I. (2011). More similar than different? Exploring cultural models of depression among Latino immigrants in Florida. *Depression Research and Treatment,* 1-11.

Meffert, S. M., & Marmar, C. R. (2009). Darfur refugees in Cairo mental health and interpersonal conflict in the aftermath of genocide. *Journal of Interpersonal Violence, 24*(11), 1835-1848.

Miller, K. E., Omidian, P., Quraishy, A. S., Quraishy, N., Nasiry, M. N., Nasiry, S., . . . Yaqubi, A. A. (2006). The afghan symptom checklist: A culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry, 76*(4), 423-433.

Mosotho, N., Louw, D., Calitz, F., & Esterhuyse, K. (2008). Depression among Sesotho speakers in Mangaung, South Africa. *African Journal of Psychiatry, 11*(1), 35-43.

Mumford, D. B., Ayub, M., Karim, R., Izhar, N., Asif, A., & Bavington, J. T. (2005). Development and validation of a questionnaire for anxiety and depression in Pakistan. *Journal of Affective Disorders, 88*(2), 175-182.

Murray, L., Bass, J., & Bolton, P. (2006). Qualitative study to identify indicators of psychological problems and functional impairment among residents of Sange district, south Kivu, eastern DRC. *A report to the Victims of Torture Fund, United States Agency for International Development (USAID).*

Naeem, F., Ayub, M., Kingdon, D., & Gobbi, M. (2012). Views of depressed patients in Pakistan concerning their illness, its causes, and treatments. *Qualitative Health Research, 22*(8), 1083-1093.

Nakimuli-Mpungu, E., Mojtabai, R., Alexandre, P. K., Katabira, E., Musisi, S., Nachega, J. B., & Bass, J. K. (2012). Cross-cultural adaptation and validation of the self-reporting questionnaire among HIV individuals in a rural ART program in southern Uganda. *HIV AIDS (Auckl), 4,* 51-60.

Nazroo, J., & O'Connor, W. (2002). *Idioms of mental distress*. London: National Centre for Social Research.

Nichter, M. (1981). Idioms of distress: Alternatives in the expression of psychosocial distress: A case study from south India. *Culture, Medicine and Psychiatry, 5*(4), 379-408.

Nieuwsma, J. A. (2011). Indigenous perspectives on depression in rural regions of India and the United States. *Transcultural Psychiatry, 48*(5), 539-568.

Núñez, L. (2009). Is it possible to eradicate poverty without attending to mental health? Listening to migrant workers in Chile through their idioms of distress. *Journal of Health Management, 11*(2), 337-354.

Okello, E. S., & Neema, S. (2007). Explanatory models and help-seeking behavior: Pathways to psychiatric care among patients admitted for depression in Mulago hospital, Kampala, Uganda. *Qualitative Health Research, 17*(1), 14-25.

Okello, E. S., Ngo, V. K., Ryan, G., Musisi, S., Akena, D., Nakasujja, N., & Wagner, G. (2012). Qualitative study of the influence of antidepressants on the psychological health of patients on antiretroviral therapy in Uganda. *African Journal of AIDS research, 11*(1), 27-44.

Parker, G., Gladstone, G., & Chee, K. T. (2001). Depression in the planet's largest ethnic group: The Chinese. *The American Journal of Psychiatry, 158*(6), 857-864.

Patel, V., Simunyu, E., Gwanzura, F., Lewis, G., & Mann, A. (1997). The Shona symptom questionnaire: The development of an indigenous measure of common mental disorders in harare. *Acta Psychiatrica Scandinavica, 95*(6), 469-475.

Patel, V. (1995). Explanatory models of mental illness in Sub-Saharan Africa. *Social Science & Medicine, 40*(9), 1291-1298.

Pereira, B., Andrew, G., Pai, R., Pelto, P., & Patel, V. (2007). The explanatory models of depression in low income countries: Listening to women in India. *Journal of Affective Disorders, 102*(1), 209-218.

Phan, T., Steel, Z., & Silove, D. (2004). An ethnographically derived measure of anxiety, depression and somatization: The Phan Vietnamese psychiatric scale. *Transcultural Psychiatry, 41*(2), 200-232.

Pincay, I. E. M., & Guarnaccia, P. J. (2007). 'It's like going through an earthquake': Anthropological perspectives on depression among latino immigrants. *Journal of Immigrant and Minority Health, 9*(1), 17-28.

Poleshuck, E. L., Cerrito, B., Leshoure, N., Finocan-Kaag, G., & Kearney, M. H. (2013). Underserved women in a women's health clinic describe their experiences of depressive symptoms and why they have low uptake of psychotherapy. *Community Mental Health Journal, 49*(1), 50-60.

Poudyal, B., Bass, J., Subyantoro, T., Jonathan, A., Erni, T., & Bolton, P. (2009). Assessment of the psychosocial and mental health needs, dysfunction and coping mechanisms of violence affected populations in Bireuen, Aceh. *Torture: Quarterly Journal on Rehabilitation of Torture Victims and Prevention of Torture, 19*(3), 218-226.

Raguram, R., Weiss, M. G., Keval, H., & Channabasavanna, S. M. (2001). Cultural dimensions of clinical depression in Bangalore, India. *Anthropology & Medicine, 8*(1), 31-46.

Rao, D., Horton, R., & Raguram, R. (2012). Gender inequality and structural violence among depressed women in south India. *Social Psychiatry and Psychiatric Epidemiology, 47*(12), 1967-1975.

Rasmussen, A., Eustache, E., Raviola, G., Kaiser, B., Grelotti, D. J., & Belkin, G. S. (2014). Development and validation of a Haitian creole screening instrument for depression. *Transcultural Psychiatry, 52*(1), 33-57.

Rees, S., & Silove, D. (2011). Sakit Hati: A state of chronic mental distress related to resentment and anger amongst West Papuan refugees exposed to persecution. *Social Science & Medicine, 73*(1), 103-110.

Rodrigues, M., Patel, V., Jaswal, S., & de Souza, N. (2003). Listening to mothers: Qualitative studies on motherhood and depression from Goa, India. *Social Science & Medicine,* 57(10), 1797-1806.

Selim, N. (2010). Cultural dimensions of depression in Bangladesh: A qualitative study in two villages of Matlab. *Journal of Health, Population, and Nutrition, 28*(1), 95-106.

Sellers, S. L., Ward, E. C., & Pate, D. (2006). Dimensions of depression: A qualitative study of wellbeing among black African immigrant women. *Qualitative Social Work: Research and Practice, 5*(1), 45-66.

Shankar, B. R., Saravanan, B., & Jacob, K. S. (2006). Explanatory models of common mental disorders among traditional healers and their patients in rural south India. *The International Journal of Social Psychiatry, 52*(3), 221-233.

Shin, J. K. (2010). Understanding the experience and manifestation of depression among korean immigrants in new york city. *Journal of Transcultural Nursing, 21*(1), 73-80.

Sin, M. K., Jordan, P., & Park, J. (2011). Perceptions of depression in Korean American immigrants. *Issues in Mental Health Nursing, 32*(3), 177-183.

Sulaiman, S. O. Y., Bhugra, D., & De Silva, P. (2001). The development of a culturally sensitive symptom checklist for depression in Dubai. *Transcult.Psychiatry, 38*(2), 219-229.

Templeton, L., Velleman, R., Persaud, A., & Milner, P. (2003). The experiences of postnatal depression in women from black and minority ethnic communities in Wiltshire, UK. *Ethnicity and Health, 8*(3), 207-221.

Tilbury, F. (2007). "I feel I am a bird without wings": Discourses of sadness and loss among east Africans in western Australia. *Identities, 14*(4), 433-458.

Ventevogel, P., Jordans, M., Reis, R., & de Jong, J. (2013). Madness or sadness? Local concepts of mental illness in four conflict-affected African communities. *Conflict and Health, 7*(1), 3-19.

Waite, R., & Killian, P. (2009). Perspectives about depression: Explanatory models among African-American women. *Archives of Psychiatric Nursing, 23*(4), 323-333.

Walters, V., Avotri, J. Y., & Charles, N. (1999). "Your heart is never free": Women in wales and Ghana talking about distress. *Canadian Psychology/Psychologie Canadienne, 40*(2), 129-142.

White, P. (2004). Heat, balance, humors, and ghosts: Postpartum in Cambodia. *Health Care for Women International, 25*(2), 179-194.

Wilk, C. M., & Bolton, P. (2002). Local perceptions of the mental health effects of the Uganda acquired immunodeficiency syndrome epidemic. *The Journal of Nervous and Mental Disease, 190*(6), 394-397.

Youngmann, R., Minuchin-Itzigsohn, S., & Barasch, M. (1999). Manifestations of emotional distress among Ethiopian immigrants in Israel: Patient and clinician perspectives. *Transcultural Psychiatry, 36*(1), 45-63.

Appendix C. Frequency of symptoms from literature review

Table A1.
Frequency (%) of DSM-5 depression symptoms across study populations

Symptoms	Study populations (*N = 138*)	Female populations (n = 49)	Male populations (n = 6)	Trauma populations (*n* = 21)	Perinatal populations (*n* = 15)
Depressed mood	93	29 (59.2)	5 (83.3)	16 (73.7)	6 (40.0)
Fatigue	91	32 (65.3)	3 (50.0)	13 (63.2)	11 (73.3)
Sleep	86	28 (57.1)	3 (50.0)	17 (84.2)	7 (46.7)
Weight/appetite	42	20 (40.8)	4 (66.7)	15 (73.7)	7 (46.7)
Thoughts of death	60	19 (38.8)	3 (50.0)	11 (52.6)	6 (40.0)
Loss of interest	59	15 (30.6)	3 (50.0)	9 (47.4)	3 (20.0)
Worthlessness/Guilt	44	13 (26.5)	1 (16.7)	9 (42.1)	4 (26.7)
Irritability	40	15 (30.6)	3 (50.0)	2 (10.5)	7 (46.7)
Concentration	33	7 (14.3)	1 (16.7)	7 (36.8)	1 (6.7)
Functioning	30	11 (22.4)	1 (16.7)	9 (42.1)	2 (13.3)
Psychomotor	15	4 (8.2)	0 (0.0)	2 (10.5)	0 (0.0)

Table A2.
Frequency (%) of DSM-5 depression symptoms across regions

Symptoms	Western non-indigenous (*n* = 29)	Western-indigenous (*n* = 2)	Latin America (*n* = 21)	Middle East/North Africa (*n* = 8)	East Asia (*n* = 11)	South Asia (*n* = 24)	Southeast Asia (*n* = 9)	Sub-Saharan Africa (*n* = 34)
Depressed mood	20 (69.0)	1 (50.0)	13 (61.9)	7 (87.5)	6 (54.5)	14 (58.3)	7 (75.0)	24 (70.6)
Fatigue	17 (58.6)	0 (0.0)	16 (76.2)	4 (50.0)	10 (90.1)	16 (66.7)	5 (62.5)	23 (67.6)
Sleep	16 (55.2)	1 (50.0)	8 (38.1)	4 (50.0)	9 (81.8)	18 (75.0)	6 (62.5)	24 (70.6)
Weight/appetite	9 (31.0)	0 (0.0)	6 (28.6)	3 (37.5)	8 (72.7)	13 (54.2)	6 (75.0)	21 (61.8)
Thoughts of death	12 (41.4)	1 (50.0)	6 (28.6)	4 (50.0)	8 (72.7)	8 (33.3)	4 (37.5)	17 (50.0)
Loss of interest	11 (37.9)	0 (0.0)	14 (66.7)	2 (25.0)	8 (72.7)	7 (29.2)	3 (37.5)	14 (41.2)
Worthlessness/Guilt	7 (24.1)	0 (0.0)	3 (14.3)	2 (25.0)	9 (81.8)	6 (25.0)	5 (62.5)	12 (35.3)
Irritability	10 (34.5)	1 (50.0)	1 (4.8)	5 (62.5)	5 (45.5)	7 (29.2)	1 (12.5)	10 (29.4)
Concentration	8 (27.6)	0 (0.0)	2 (9.5)	0 (0.0)	7 (63.6)	5 (20.8)	3 (37.5)	8 (23.5)
Functioning	6 (20.7)	0 (0.0)	1 (4.8)	2 (25.0)	1 (9.1)	8 (33.3)	1 (12.5)	11 (32.4)
Psychomotor	3 (10.3)	0 (0.0)	2 (9.5)	0 (0.0)	3 (27.3)	2 (8.3)	1 (12.5)	4 (11.8)

Table A3.
Frequency (%) of non-DSM-5 depression symptoms across study populations

Symptoms	Study populations (*N = 138*)	Female populations (n = 49)	Male populations (n = 6)	Trauma populations (*n* = 19)	Perinatal populations (*n* = 15)
Social isolation/loneliness	68 (49.3)	20 (40.8)	5 (83.3)	15 (78.9)	8 (53.3)
Crying	64 (46.3)	19 (38.8)	3 (50.0)	10 (52.6)	8 (53.3)
General pain	53 (38.4)	19 (38.8)	1 (16.7)	5 (26.3)	2 (13.3)
Anger	50 (36.2)	13 (26.5)	2 (33.3)	9 (47.4)	7 (46.7)
Headache	50 (36.2)	19 (38.8)	1 (16.7)	4 (21.1)	2 (13.3)
Heart issues	46 (33.3)	15 (30.6)	0 (0.0)	7 (36.8)	4 (26.7)
Thinking too much	44 (31.9)	16 (32.7)	3 (50.0)	9 (47.4)	5 (33.3)
Hopelessness	41 (29.7)	11 (22.4)	1 (16.7)	12 (63.2)	3 (20.0)
Worry	39 (28.3)	14 (28.6)	2 (33.3)	1 (5.3)	5 (33.3)
Stomach aches	28 (20.3)	13 (26.5)	2 (33.3)	0 (0.0)	2 (13.3)
Weakness	27 (19.6)	9 (18.4)	1 (16.7)	6 (31.6)	2 (13.3)
Anxiety	28 (20.3)	10 (20.4)	0 (0.0)	1 (5.3)	4 (26.7)
Nervousness/tense	26 (18.8)	7 (14.3)	1 (16.7)	5 (26.3)	1 (6.7)
Confusion	25 (18.1)	5 (10.2)	1 (16.7)	8 (42.1)	2 (13.3)
Stressed	25 (18.1)	9 (18.4)	1 (16.7)	3 (15.8)	4 (26.7)
Scared	23 (16.7)	8 (16.3)	0 (0.0)	2 (10.5)	4 (26.7)
Self-esteem	19 (13.8)	6 (12.2)	1 (16.7)	2 (10.5)	3 (20.0)
Memory	18 (13.0)	5 (10.2)	0 (0.0)	6 (31.6)	0 (0.0)
Breathing issues	19 (13.8)	8 (16.3)	0 (0.0)	4 (21.1)	1 (6.7)
Emptiness	18 (13.0)	8 (16.3)	1 (16.7)	1 (5.3)	3 (20.0)
Grooming problems	18 (13.0)	4 (8.2)	2 (33.3)	4 (21.1)	1 (6.7)
Not talking/talking all the time	19 (13.8)	3 (6.1)	2 (33.3)	5 (26.3)	0 (0.0)
Dizziness	17 (12.3)	5 (10.2)	0 (0.0)	1 (5.3)	3 (20.0)
Unhealthy	17 (12.3)	3 (6.1)	1 (16.7)	6 (31.6)	2 (13.3)
Head issues	17 (12.3)	3 (6.1)	0 (0.0)	4 (21.1)	0 (0.0)
Hot and cold sensations	16 (11.6)	3 (6.1)	1 (16.7)	2 (10.5)	0 (0.0)
Frustration	16 (11.6)	5 (10.2)	0 (0.0)	2 (10.5)	3 (20.0)
Sad appearance	16 (11.6)	7 (14.3)	1 (16.7)	1 (5.3)	0 (0.0)
Substance use/abuse	16 (11.6)	2 (4.1)	0 (0.0)	3 (15.8)	0 (0.0)
Loss of control	14 (10.1)	3 (6.1)	1 (16.7)	2 (10.5)	1 (6.7)
Interpersonal problems	14 (10.1)	5 (10.2)	1 (16.7)	3 (15.8)	4 (26.7)
Suspicious	13 (9.4)	2 (4.1)	1 (16.7)	8 (42.1)	0 (0.0)
Restless	12 (8.7)	5 (10.2)	0 (0.0)	4 (21.1)	0 (0.0)
Rumination	10 (7.3)	2 (4.1)	0 (0.0)	7 (36.8)	0 (0.0)
Unstable	12 (8.7)	3 (6.1)	0 (0.0)	3 (15.8)	0 (0.0)
Nausea	10 (7.3)	3 (6.1)	0 (0.0)	3 (15.8)	0 (0.0)
Disappointed	11 (8.0)	5 (10.2)	0 (0.0)	4 (21.1)	3 (20.0)
Aggression	9 (6.5)	2 (4.1)	1 (16.7)	1 (5.3)	1 (6.7)
Change from before	10 (7.3)	1 (2.0)	2 (33.3)	0 (0.0)	1 (6.7)

Feeling dead	8 (5.8)	2 (4.1)	0 (0.0)	3 (15.8)	0 (0.0)
Blackness	9 (6.5)	5 (10.2)	0 (0.0)	0 (0.0)	1 (6.7)
Embarrassed	9 (6.5)	3 (6.1)	0 (0.0)	0 (0.0)	2 (13.3)
Trapped	9 (6.5)	7 (14.3)	0 (0.0)	0 (0.0)	1 (6.7)
Trembling	9 (6.5)	1 (2.0)	0 (0.0)	2 (10.5)	0 (0.0)
Upset	9 (6.5)	2 (4.1)	0 (0.0)	1 (5.3)	1 (6.7)
Nightmares	8 (5.8)	1 (2.0)	0 (0.0)	4 (21.1)	0 (0.0)
Pessimism	8 (5.8)	2 (4.1)	2 (33.3)	0 (0.0)	0 (0.0)
Psychotic symptoms	8 (5.8)	3 (6.1)	0 (0.0)	0 (0.0)	1 (6.7)
Staying in bed	9 (6.5)	3 (6.1)	2 (33.3)	2 (10.5)	1 (6.7)
Despair	7 (5.1)	2 (4.1)	0 (0.0)	0 (0.0)	0 (0.0)
Chest pressure/tightness	6 (4.3)	3 (6.1)	0 (0.0)	0 (0.0)	0 (0.0)
Constipation	6 (4.3)	1 (2.0)	0 (0.0)	0 (0.0)	0 (0.0)
Harming others	6 (4.3)	5 (10.2)	0 (0.0)	1 (5.3)	2 (13.3)
Heavy body	6 (4.3)	4 (8.2)	0 (0.0)	0 (0.0)	0 (0.0)
Regretful	6 (4.3)	1 (2.0)	0 (0.0)	4 (21.1)	2 (13.3)
Digestion	5 (3.6)	1 (2.0)	0 (0.0)	3 (15.8)	0 (0.0)
Eye problems	5 (3.6)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Desperation	5 (3.6)	0 (0.0)	0 (0.0)	1 (5.3)	0 (0.0)
Feeling faint	5 (3.6)	2 (4.1)	0 (0.0)	2 (10.5)	1 (6.7)
Genital complaints	5 (3.6)	4 (8.2)	0 (0.0)	0 (0.0)	0 (0.0)
Grief	6 (3.6)	2 (4.1)	0 (0.0)	2 (10.5)	0 (0.0)
Loss of libido	5 (3.6)	3 (6.1)	0 (0.0)	0 (0.0)	1 (6.7)
Bad behavior	4 (2.9)	0 (0.0)	1 (16.7)	1 (5.3)	0 (0.0)
Helpless	4 (2.9)	2 (4.1)	1 (16.7)	0 (0.0)	0 (0.0)
Bored	4 (2.9)	2 (4.1)	0 (0.0)	1 (5.3)	0 (0.0)
Disputing/arguing	4 (2.9)	1 (2.0)	0 (0.0)	2 (10.5)	1 (6.7)
Fever	4 (2.9)	1 (2.0)	0 (0.0)	0 (0.0)	1 (6.7)
Inability to move	4 (2.9)	1 (2.0)	0 (0.0)	1 (5.3)	0 (0.0)
Falling	3 (2.2)	0 (0.0)	0 (0.0)	2 (10.5)	0 (0.0)
Lack of coping	3 (2.2)	2 (4.1)	0 (0.0)	0 (0.0)	1 (6.7)
Lazy	3 (2.2)	1 (2.0)	0 (0.0)	1 (5.3)	0 (0.0)
Startled	2 (1.4)	0 (0.0)	0 (0.0)	1 (5.3)	0 (0.0)
Useless	3 (2.2)	1 (2.0)	0 (0.0)	0 (0.0)	0 (0.0)
Homesickness	2 (1.4)	1 (2.0)	1 (16.7)	0 (0.0)	0 (0.0)
Lack of peace	2 (1.4)	1 (2.0)	0 (0.0)	0 (0.0)	1 (6.7)
Panic	2 (1.4)	1 (2.0)	0 (0.0)	0 (0.0)	1 (6.7)
Self-harm	2 (1.4)	1 (2.0)	0 (0.0)	1 (5.3)	0 (0.0)
Sweating a lot	2 (1.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Unable to laugh	2 (1.4)	1 (2.0)	0 (0.0)	0 (0.0)	1 (6.7)

Table A4.
Frequency (%) of non-DSM-5 depression symptoms across study populations

Symptoms	Western non-indigenous (*n* = 29)	Western-indigenous (*n* = 2)	Latin America (*n* = 21)	Middle East/North Africa (*n* = 8)	East Asia (*n* = 11)	South Asia (*n* = 24)	Southeast Asia (*n* = 9)	Sub-Saharan Africa (*n* = 34)
Social isolation/loneliness	17 (58.6)	1 (50.0)	11 (52.4)	4 (50.0)	3 (27.3)	8 (33.3)	6 (75.0)	18 (52.9)
Crying	12 (41.4)	0 (0.0)	14 (66.7)	5 (62.5)	5 (45.5)	11 (45.8)	3 (37.5)	13 (38.2)
General pain	6 (20.7)	1 (50.0)	8 (38.1)	4 (50.0)	6 (54.6)	10 (41.7)	4 (50.0)	14 (41.1)
Anger	11 (37.9)	0 (0.0)	11 (52.4)	2 (25.0)	4 (36.4)	7 (29.2)	3 (37.5)	11 (32.4)
Headache	5 (17.2)	1 (50.0)	9 (42.9)	1 (12.5)	4 (36.4)	11 (45.8)	3 (25.0)	17 (50.0)
Heart issues	6 (20.7)	1 (50.0)	5 (23.8)	2 (25.0)	4 (36.4)	11 (45.8)	7 (87.5)	11 (32.4)
Thinking too much	5 (17.2)	2 (100.0)	4 (19.0)	3 (37.5)	1 (9.1)	9 (37.5)	3 (37.5)	17 (50.0)
Hopelessness	7 (24.1)	0 (0.0)	3 (14.3)	2 (25.0)	5 (45.5)	6 (25.0)	4 (50.0)	14 (38.2)
Worry	8 (27.6)	1 (50.0)	3 (14.3)	0 (0.0)	1 (9.1)	9 (37.5)	3 (37.5)	13 (38.2)
Stomach aches	3 (10.3)	0 (0.0)	2 (9.5)	3 (37.5)	1 (9.1)	4 (16.7)	0 (0.0)	15 (44.1)
Weakness	2 (6.9)	0 (0.0)	3 (14.3)	1 (12.5)	2 (18.2)	8 (33.3)	4 (50.0)	6 (17.7)
Anxiety	9 (31.0)	0 (0.0)	4 (19.0)	2 (25.0)	4 (36.4)	6 (25.0)	2 (25.0)	1 (2.9)
Nervousness/tense	6 (20.7)	0 (0.0)	7 (33.3)	3 (37.5)	1 (9.1)	5 (20.8)	2 (25.0)	2 (5.9)
Confusion	1 (3.4)	1 (50.0)	6 (28.6)	0 (0.0)	3 (27.3)	2 (8.3)	5 (62.5)	7 (20.6)
Stressed	11 (37.9)	0 (0.0)	0 (0.0)	1 (12.5)	1 (9.1)	3 (12.5)	3 (37.5)	6 (17.7)
Scared	3 (10.3)	0 (0.0)	5 (23.8)	3 (37.5)	0 (0.0)	3 (12.5)	2 (25.0)	5 (14.7)
Self-esteem	3 (10.3)	0 (0.0)	5 (23.8)	1 (12.5)	3 (27.3)	3 (12.5)	3 (37.5)	1 (2.9)
Memory	1 (3.4)	0 (0.0)	3 (14.3)	0 (0.0)	1 (9.1)	5 (20.8)	1 (12.5)	7 (20.6)
Breathing issues	1 (3.4)	0 (0.0)	0 (0.0)	4 (50.0)	4 (36.4)	5 (20.8)	2 (25.0)	3 (8.8)
Emptiness	6 (20.7)	1 (50.0)	3 (14.3)	0 (0.0)	2 (18.2)	4 (16.7)	1 (12.5)	1 (2.9)
Grooming problems	2 (6.9)	0 (0.0)	9 (42.9)	1 (12.5)	0 (0.0)	0 (0.0)	0 (0.0)	6 (17.7)
Not talking/talking all the time	3 (10.3)	0 (0.0)	4 (19.0)	1 (12.5)	0 (0.0)	2 (8.3)	2 (12.5)	7 (20.6)
Dizziness	0 (0.0)	1 (50.0)	3 (14.3)	0 (0.0)	2 (18.2)	4 (16.7)	2 (25.0)	5 (14.7)

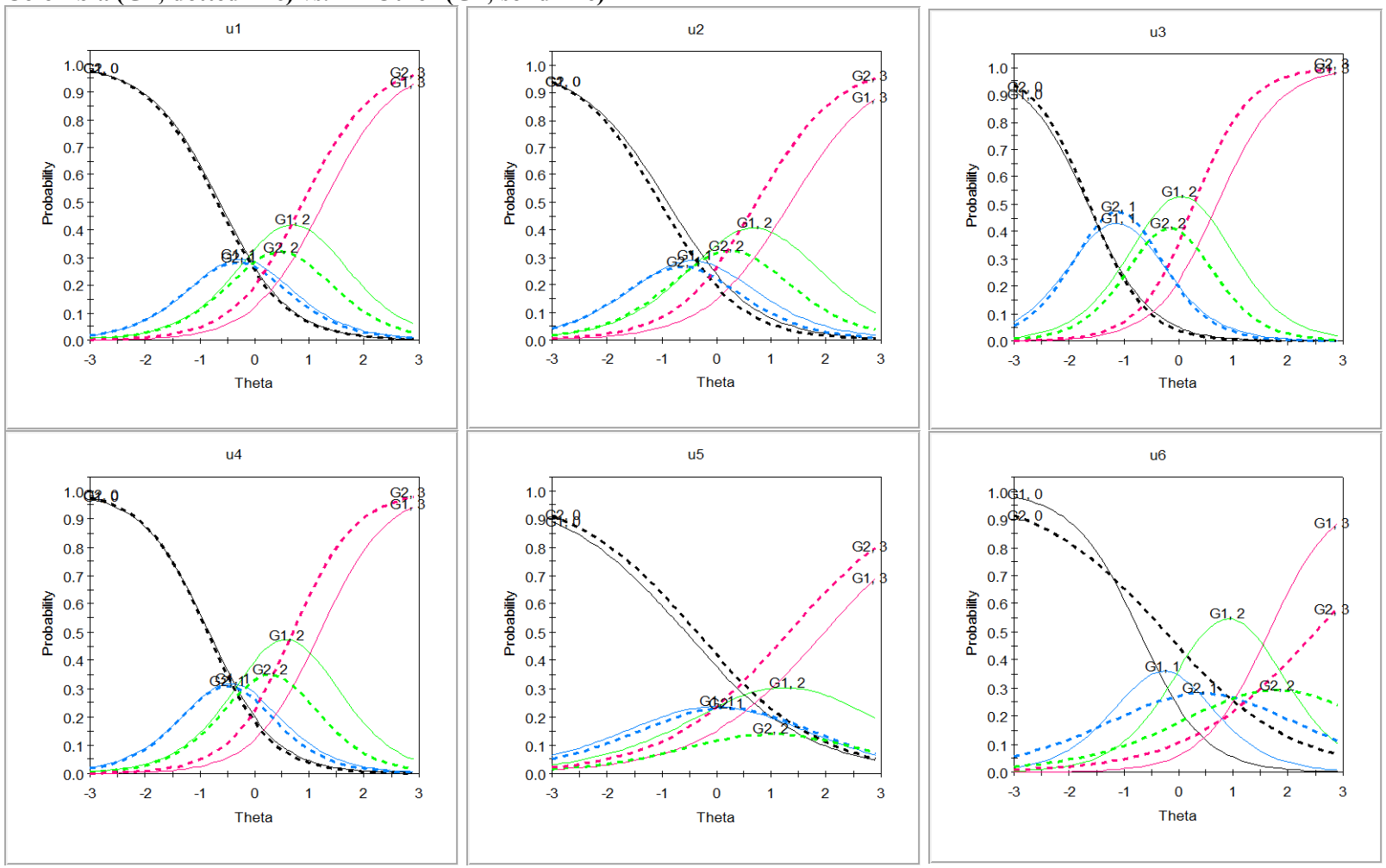
Unhealthy	1 (3.4)	1 (50.0)	2 (9.5)	0 (0.0)	4 (36.4)	2 (8.3)	4 (50.0)	3 (8.8)
Head issues	2 (6.9)	0 (0.0)	2 (9.5)	0 (0.0)	5 (45.5)	2 (8.3)	3 (37.5)	3 (8.8)
Hot and cold sensations	2 (6.9)	1 (50.0)	1 (4.8)	0 (0.0)	2 (18.2)	4 (16.7)	2 (25.0)	4 (11.8)
Frustration	4 (13.8)	0 (0.0)	2 (9.5)	0 (0.0)	1 (9.1)	6 (25.0)	1 (12.5)	2 (5.9)
Sad appearance	1 (3.4)	0 (0.0)	1 (4.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (2.9)
Substance use/abuse	2 (6.9)	1 (50.0)	10 (47.6)	0 (0.0)	0 (0.0)	2 (8.3)	0 (0.0)	1 (2.9)
Loss of control	5 (17.2)	0 (0.0)	3 (14.3)	0 (0.0)	0 (0.0)	4 (16.7)	1 (12.5)	1 (2.9)
Interpersonal problems	3 (10.3)	1 (50.0)	0 (0.0)	1 (12.5)	2 (18.2)	2 (8.3)	0 (0.0)	1 (2.9)
Suspicious	3 (10.3)	0 (0.0)	0 (0.0)	1 (12.5)	1 (9.1)	0 (0.0)	3 (37.5)	5 (14.7)
Restless	3 (10.3)	0 (0.0)	1 (4.8)	2 (25.0)	1 (9.1)	1 (4.2)	3 (37.5)	1 (2.9)
Rumination	1 (3.4)	0 (0.0)	0 (0.0)	1 (12.5)	0 (0.0)	3 (12.5)	1 (12.5)	4 (11.8)
Unstable	2 (6.9)	0 (0.0)	2 (9.5)	0 (0.0)	0 (0.0)	4 (16.7)	0 (0.0)	3 (8.8)
Nausea	1 (3.4)	0 (0.0)	1 (4.8)	1 (12.5)	2 (18.2)	1 (4.2)	3 (37.5)	1 (2.9)
Disappointed	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	3 (27.3)	2 (8.3)	5 (50.0)	0 (0.0)
Aggression	2 (6.9)	0 (0.0)	0 (0.0)	1 (12.5)	0 (0.0)	1 (4.2)	0 (0.0)	5 (14.7)
Change from before	2 (6.9)	0 (0.0)	4 (19.0)	1 (12.5)	0 (0.0)	1 (4.2)	0 (0.0)	2 (2.9)
Feeling dead	2 (6.9)	0 (0.0)	1 (4.8)	2 (25.0)	0 (0.0)	0 (0.0)	1 (12.5)	2 (5.9)
Blackness	5 (17.2)	0 (0.0)	0 (0.0)	2 (25.0)	2 (18.2)	0 (0.0)	0 (0.0)	0 (0.0)
Embarrassed	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (25.0)	6 (17.6)
Trapped	2 (6.9)	0 (0.0)	1 (4.8)	1 (12.5)	1 (9.1)	2 (8.3)	1 (12.5)	1 (2.9)
Trembling	1 (3.4)	0 (0.0)	1 (4.8)	0 (0.0)	1 (9.1)	2 (8.3)	1 (12.5)	3 (8.8)
Upset	1 (3.4)	0 (0.0)	1 (4.8)	0 (0.0)	1 (9.1)	4 (16.7)	1 (12.5)	1 (2.9)
Nightmares	0 (0.0)	1 (50.0)	2 (9.5)	1 (12.5)	1 (9.1)	1 (4.2)	0 (0.0)	2 (5.9)
Pessimism	4 (13.8)	0 (0.0)	1 (4.8)	1 (12.5)	1 (9.1)	1 (4.2)	0 (0.0)	0 (0.0)
Psychotic symptoms	0 (0.0)	0 (0.0)	0 (0.0)	2 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)	6 (17.6)
Staying in bed	4 (13.8)	0 (0.0)	1 (4.8)	0 (0.0)	1 (9.1)	0 (0.0)	0 (0.0)	3 (5.9)
Despair	2 (6.9)	0 (0.0)	2 (9.5)	0 (0.0)	0 (0.0)	2 (8.3)	1 (12.5)	0 (0.0)
Chest pressure/tightness	0 (0.0)	0 (0.0)	0 (0.0)	3 (37.5)	3 (27.3)	0 (0.0)	0 (0.0)	0 (0.0)

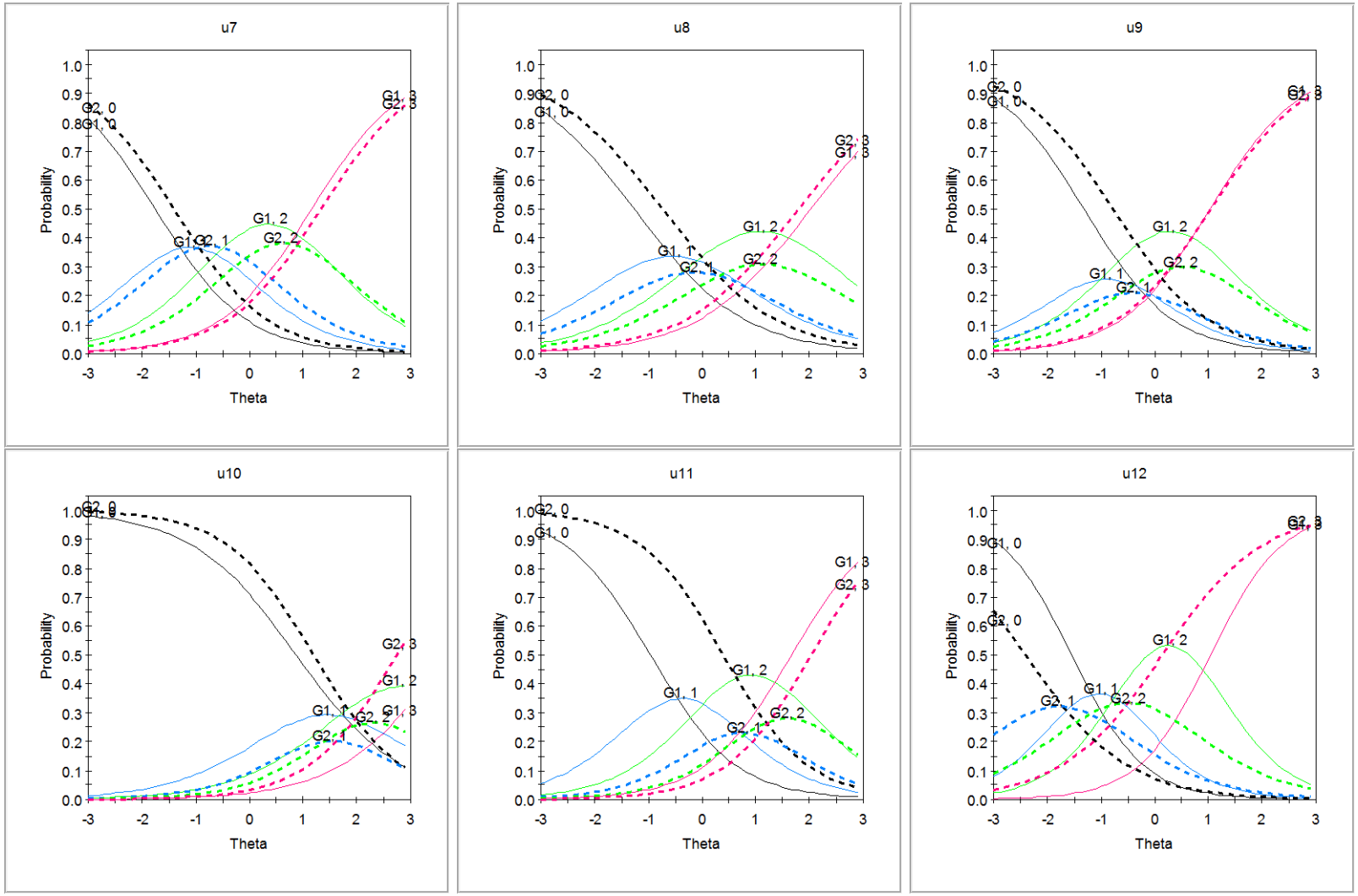
Constipation	0 (0.0)	0 (0.0)	1 (4.8)	0 (0.0)	1 (9.1)	0 (0.0)	0 (0.0)	4 (11.8)
Harming others	2 (6.9)	0 (0.0)	0 (0.0)	1 (12.5)	0 (0.0)	2 (8.3)	0 (0.0)	1 (2.9)
Heavy body	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	1 (9.1)	1 (4.2)	0 (0.0)	3 (8.8)
Regretful	1 (3.4)	0 (0.0)	0 (0.0)	1 (12.5)	0 (0.0)	0 (0.0)	1 (12.5)	3 (8.8)
Digestion	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (9.1)	1 (4.2)	3 (37.5)	0 (0.0)
Eye problems	1 (3.4)	1 (50.0)	0 (0.0)	0 (0.0)	2 (18.2)	0 (0.0)	1 (12.5)	0 (0.0)
Desperation	0 (0.0)	0 (0.0)	4 (19.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Feeling faint	1 (3.4)	0 (0.0)	1 (4.8)	1 (12.5)	0 (0.0)	2 (8.3)	0 (0.0)	0 (0.0)
Genital complaints	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	4 (16.7)	0 (0.0)	1 (2.9)
Grief	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (4.2)	2 (25.0)	3 (5.9)
Loss of libido	0 (0.0)	0 (0.0)	1 (4.8)	1 (12.5)	1 (9.1)	0 (0.0)	0 (0.0)	2 (5.9)
Bad behavior	1 (3.4)	0 (0.0)	1 (4.8)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (5.9)
Helpless	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	1 (9.1)	2 (8.3)	0 (0.0)	0 (0.0)
Bored	0 (0.0)	0 (0.0)	0 (0.0)	1 (12.5)	1 (9.1)	2 (8.3)	0 (0.0)	0 (0.0)
Disputing/arguing	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (4.2)	0 (0.0)	3 (8.8)
Fever	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	3 (8.8)
Inability to move	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (4.2)	0 (0.0)	3 (8.8)
Falling	1 (3.4)	0 (0.0)	1 (4.8)	0 (0.0)	0 (0.0)	1 (4.2)	0 (0.0)	0 (0.0)
Lack of coping	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (9.1)	0 (0.0)	0 (0.0)	0 (0.0)
Lazy	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (8.3)	0 (0.0)	0 (0.0)
Startled	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (5.9)
Useless	0 (0.0)	0 (0.0)	1 (4.8)	0 (0.0)	1 (9.1)	1 (4.2)	0 (0.0)	0 (0.0)
Homesickness	1 (3.4)	1 (50.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Lack of peace	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (4.2)	0 (0.0)	1 (2.9)
Panic	1 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (2.9)
Self-harm	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (4.2)	1 (12.5)	0 (0.0)
Sweating a lot	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	2 (5.9)
Unable to laugh	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (9.1)	0 (0.0)	0 (0.0)	1 (2.9)

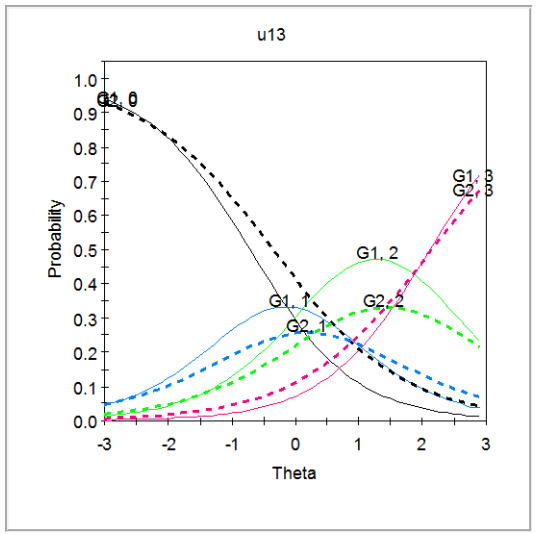
Appendix D. Item characteristic and information curves by country

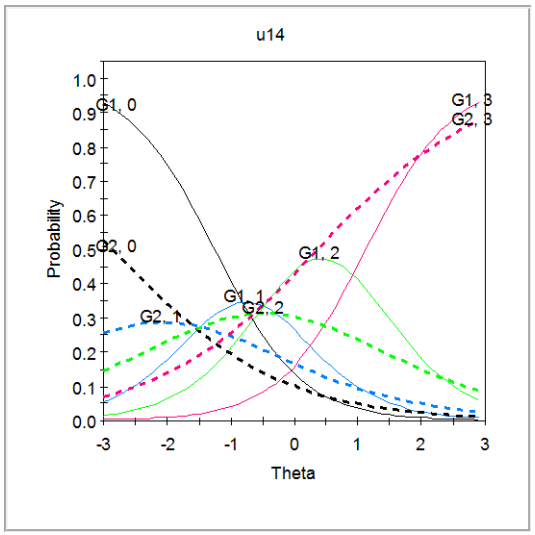
Colombia

Colombia (G2; dotted line) vs. All Other (G1; solid line)




u7
u8
u9
u10
u11
u12
Probability
Theta
G1, 0
G2, 0
G1, 1
G2, 1
G1, 2
G2, 2
G1, 3
G2, 3


u13
Probability
Theta
G1, 0
G1, 1
G1, 2
G1, 3
G2, 1
G2, 2
G2, 3


u14
Probability
Theta
G1, 0
G1, 1
G1, 2
G1, 3
G2, 0
G2, 1
G2, 2
G2, 3

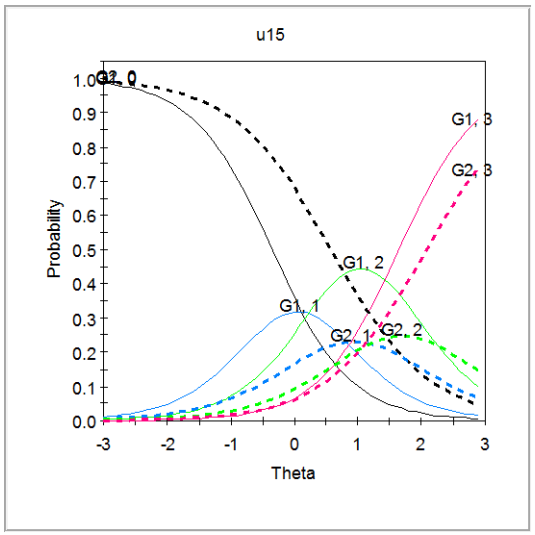
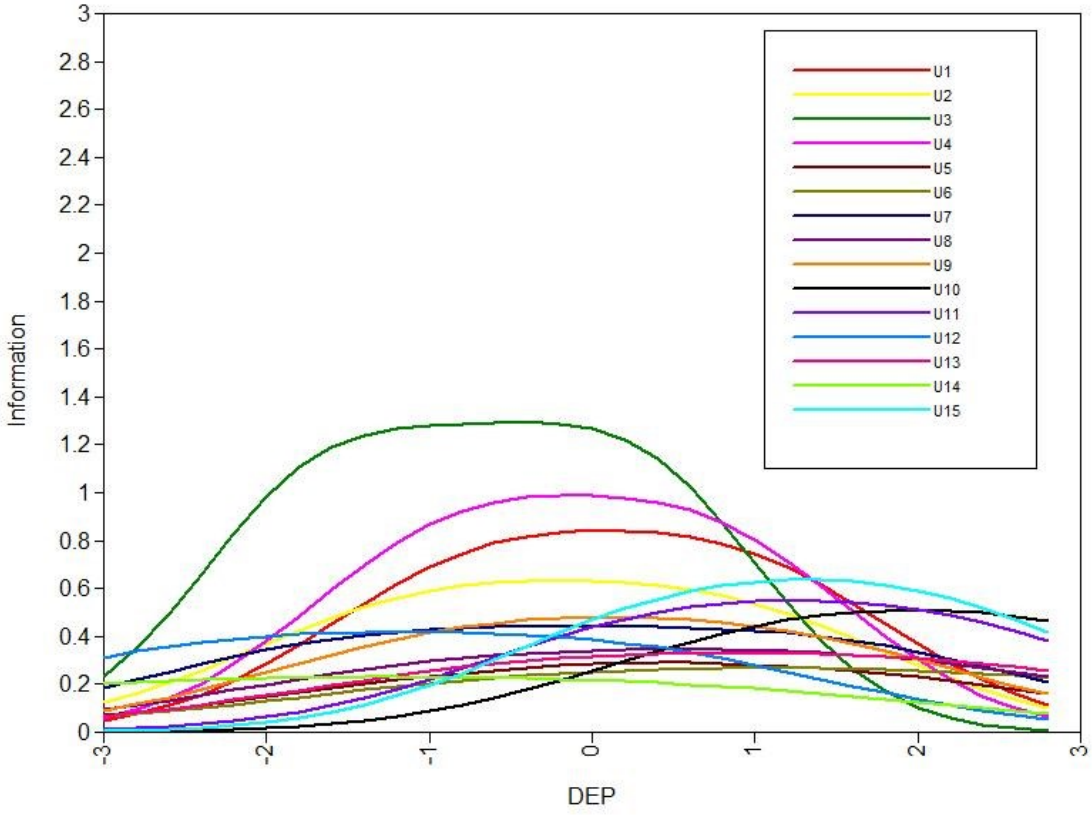
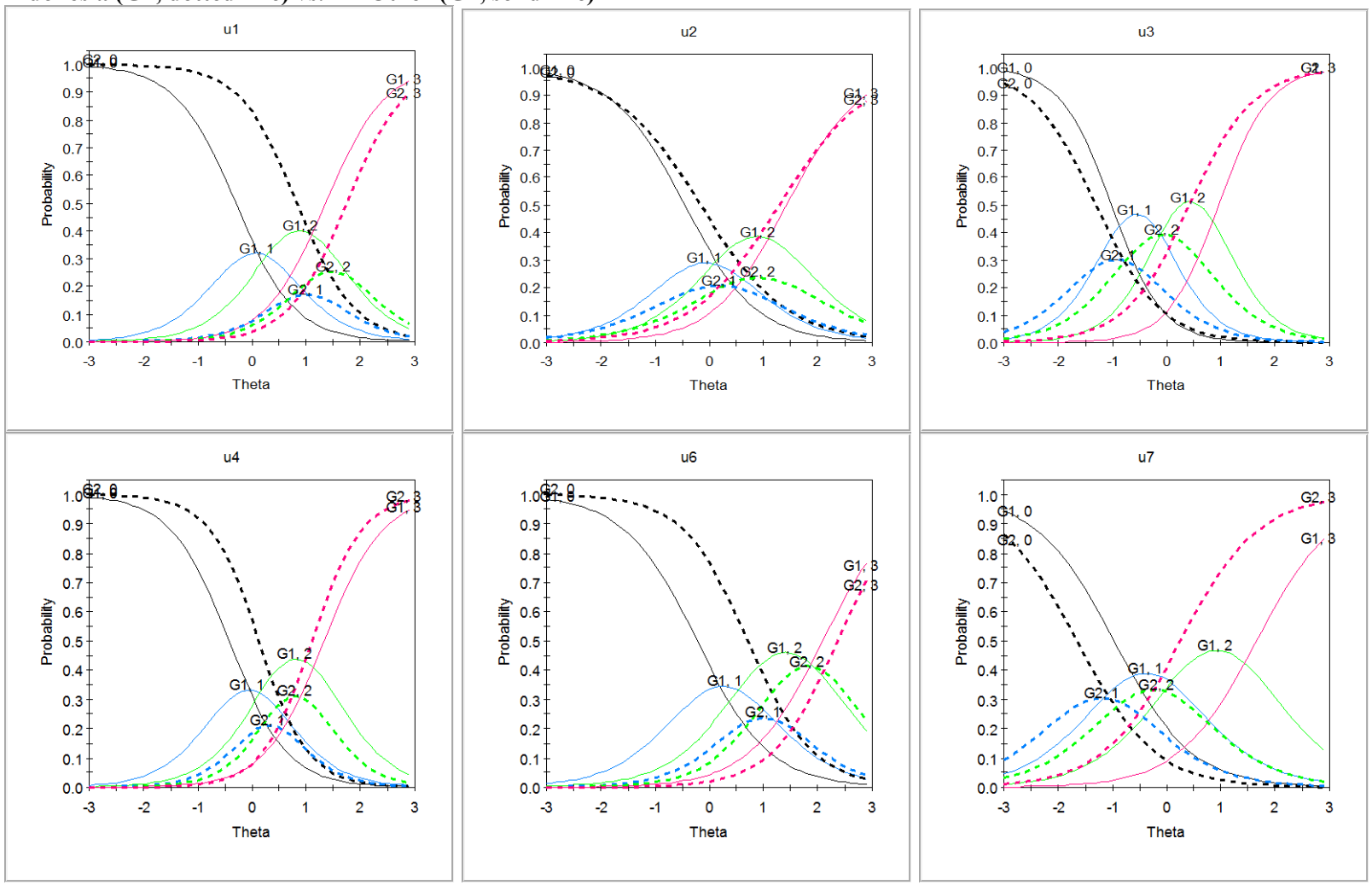

u15
Probability
Theta
G1, 1
G1, 2
G1, 3
G2, 1
G2, 2
G2, 3

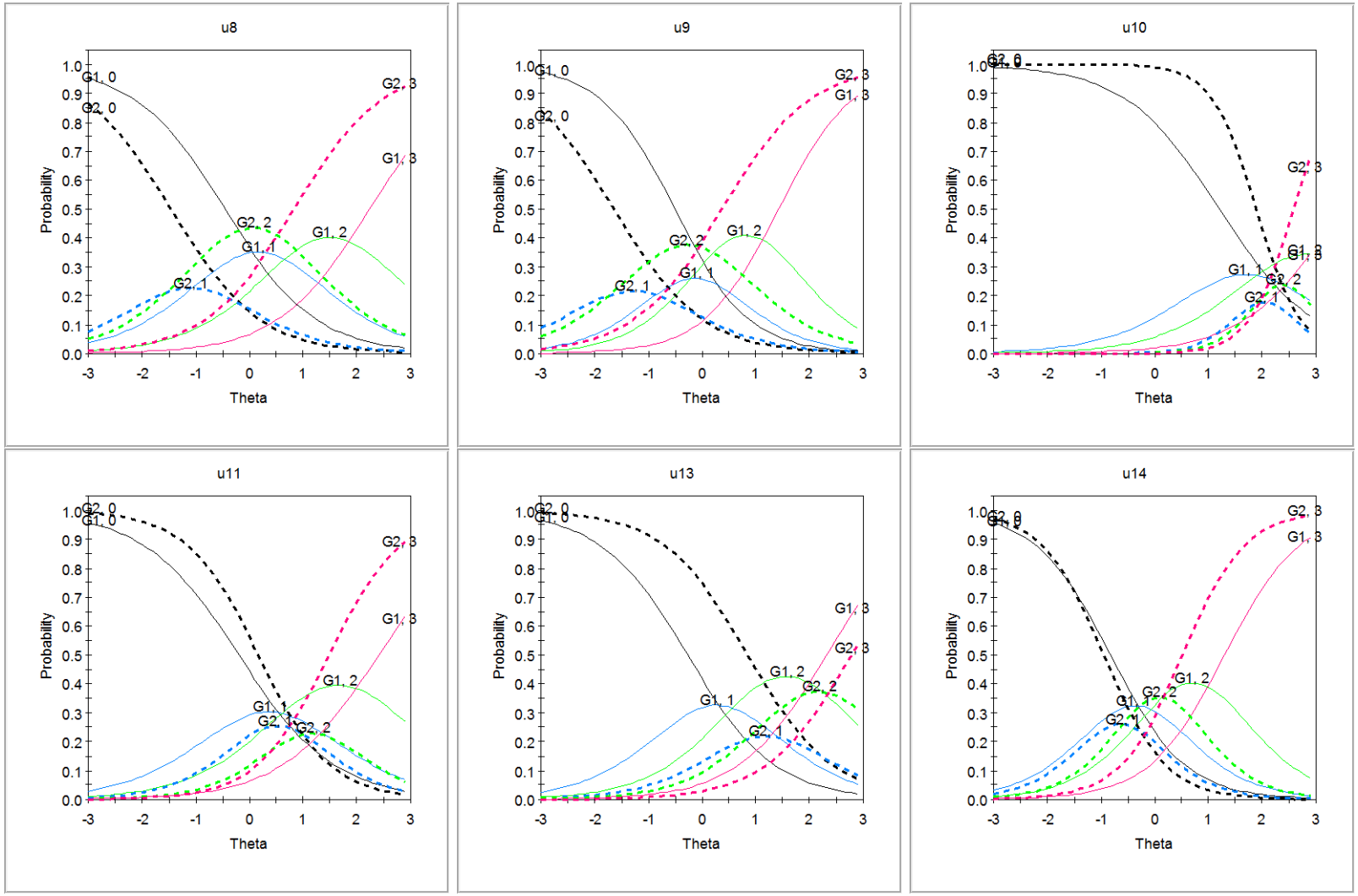
Figure A1.
Information curves for all items in Colombia



Indonesia

Indonesia (G2, dotted line) vs. All Other (G1, solid line)




u8
u9
u10
u11
u13
u14
Probability
Theta
G1, 0
G2, 0
G2, 3
G1, 3
G2, 2
G1, 2
G1, 1
G2, 1

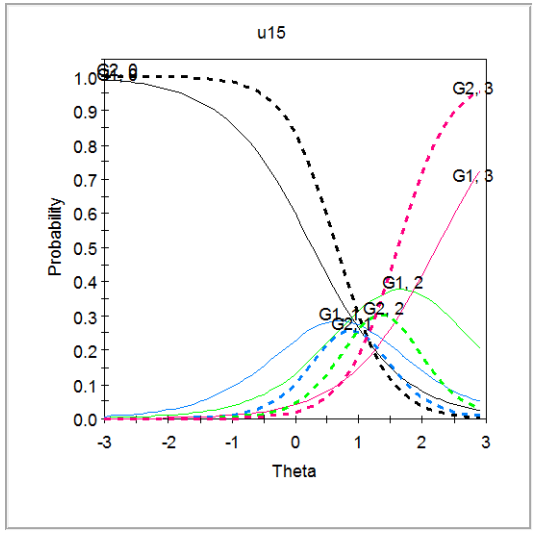
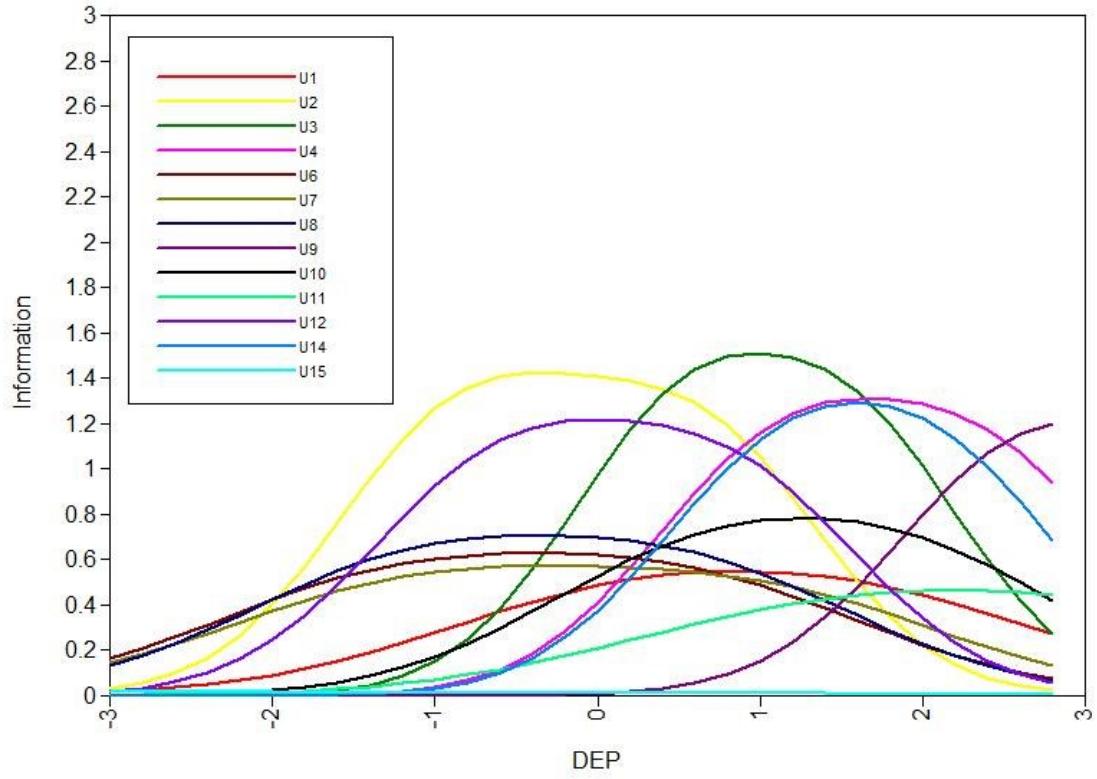
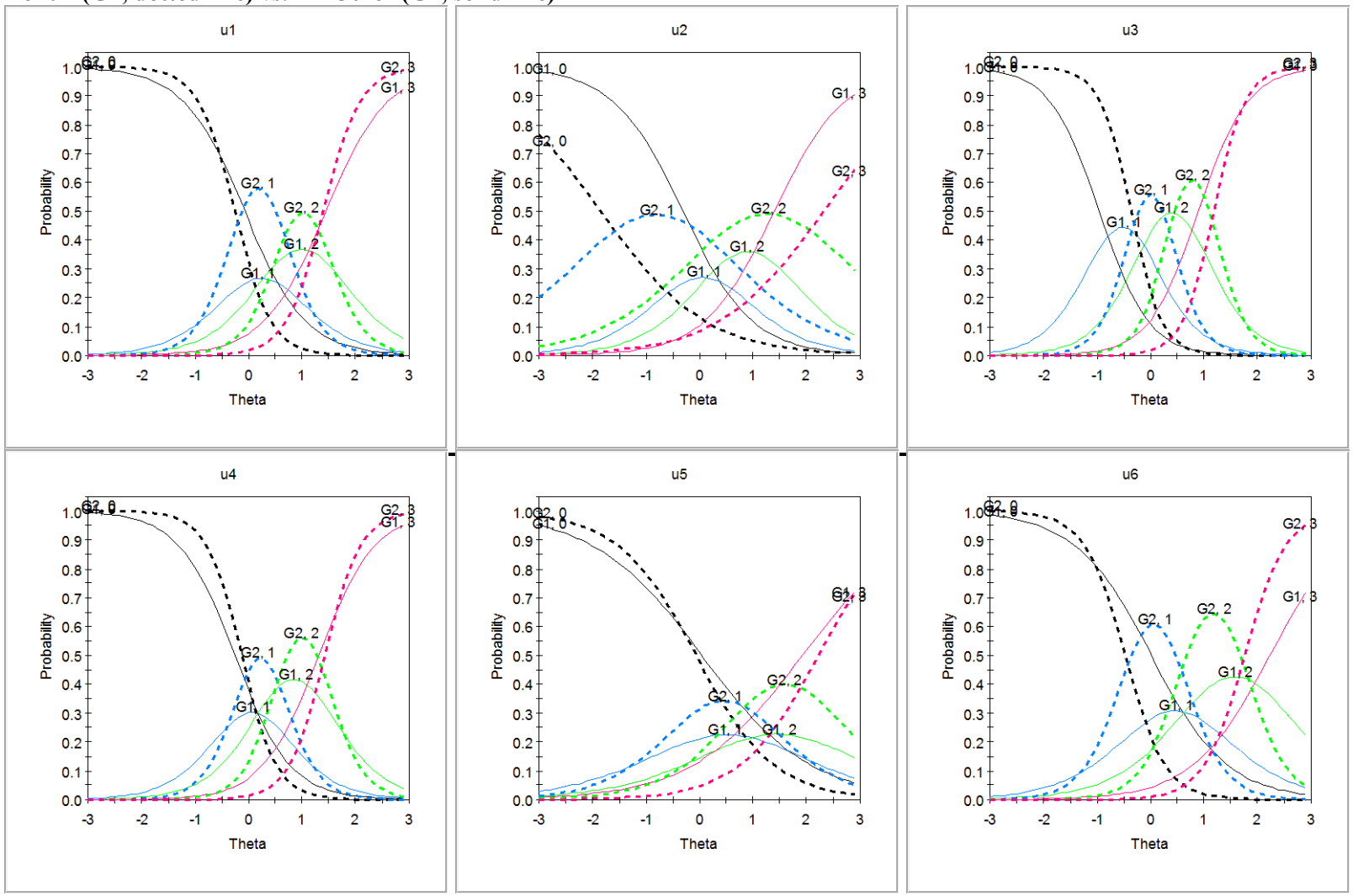

u15
Probability
Theta
G2, 3
G1, 3
G1, 2
G2, 2
G1, 1
G2, 1

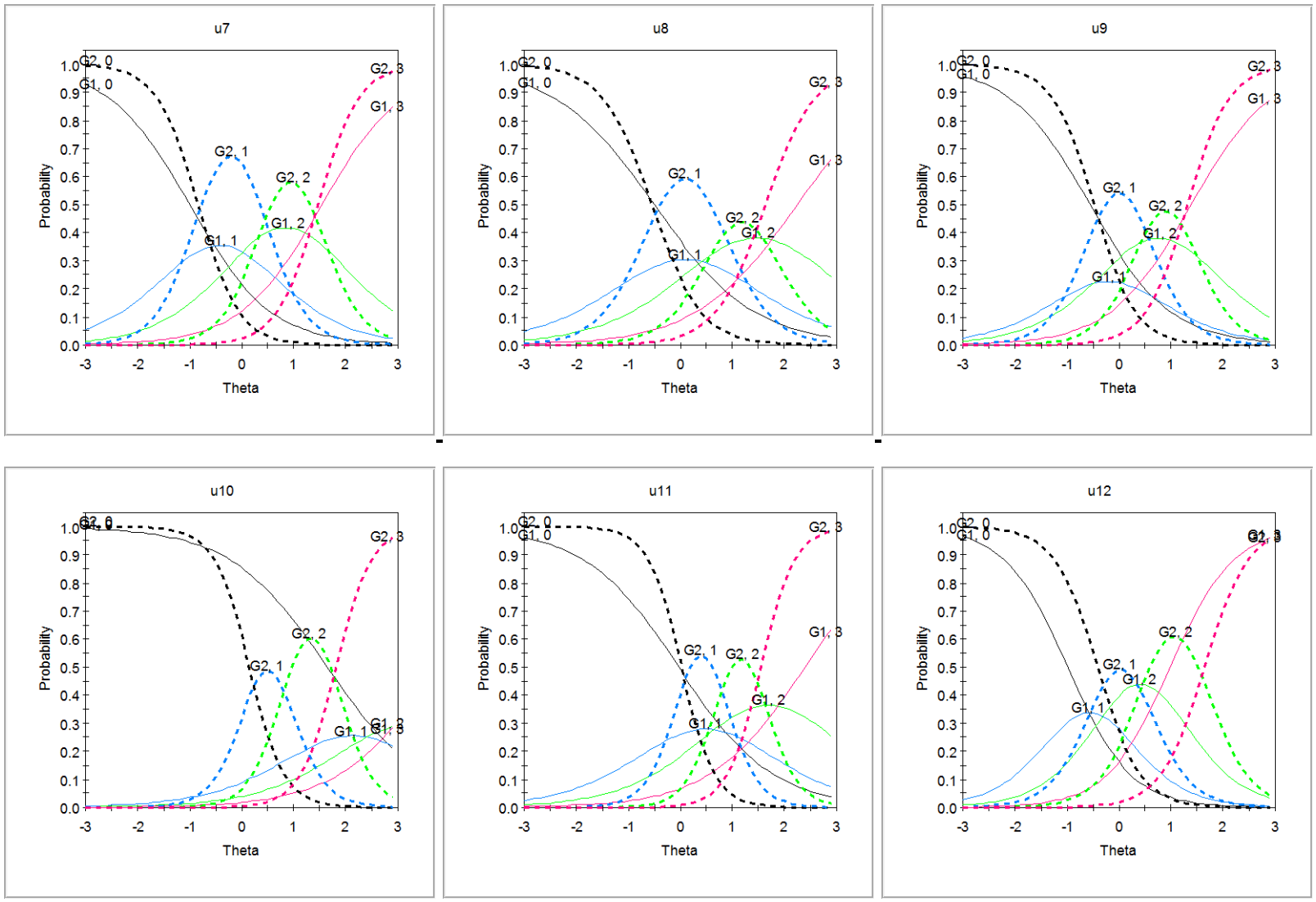
Figure A2.
Information curves for all items in Indonesia



Dohuk

Dohuk (G2, dotted line) vs. All Other (G1, solid line)




u7
u8
u9
u10
u11
u12
Probability
Theta
G2, 0
G1, 0
G2, 1
G1, 1
G2, 2
G1, 2
G2, 3
G1, 3

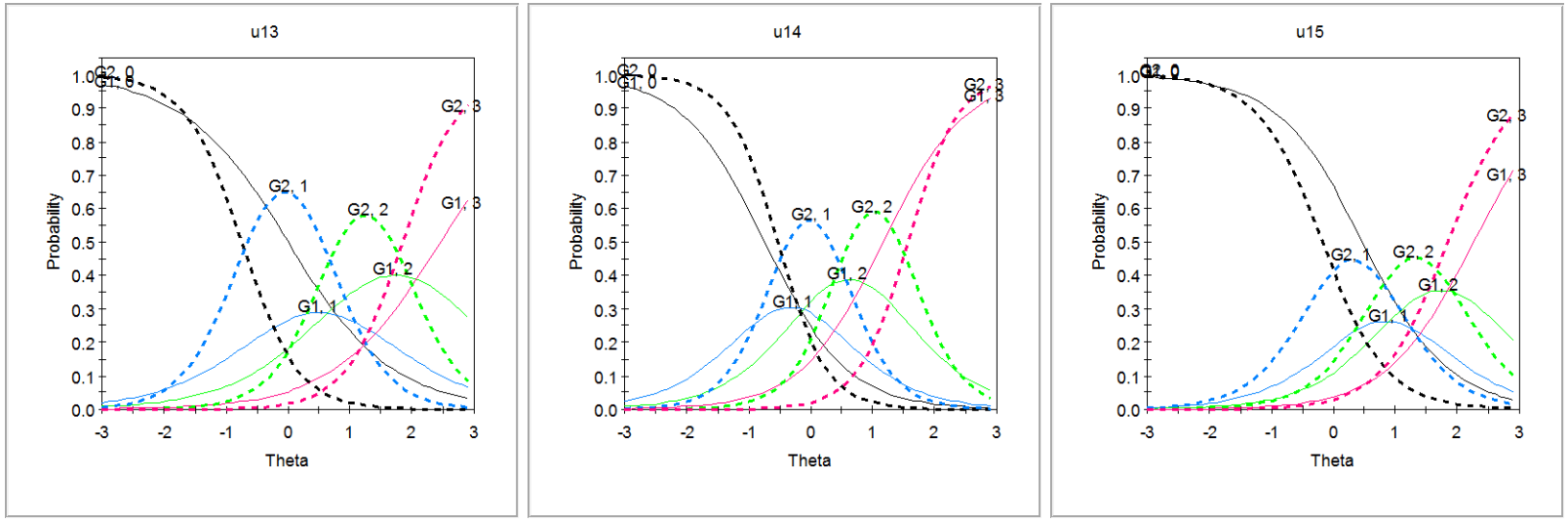
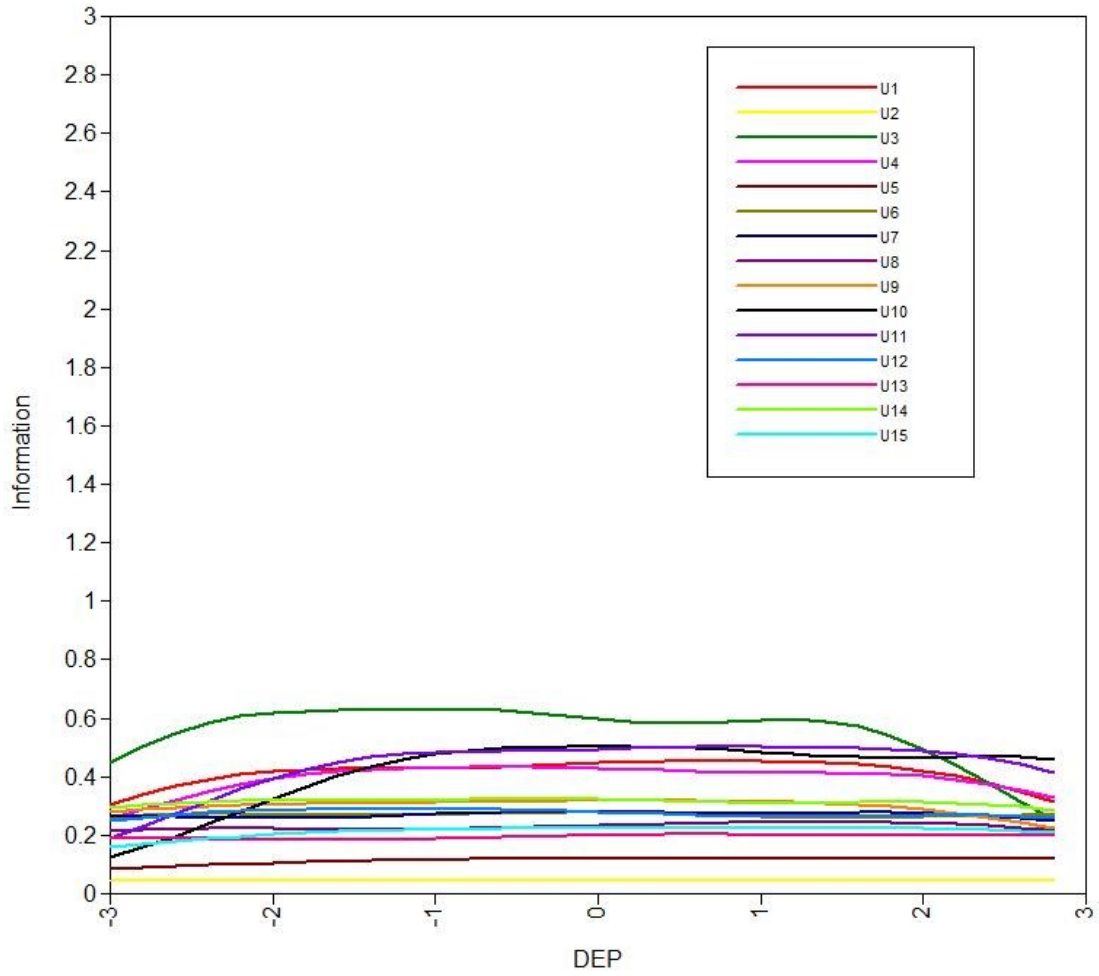
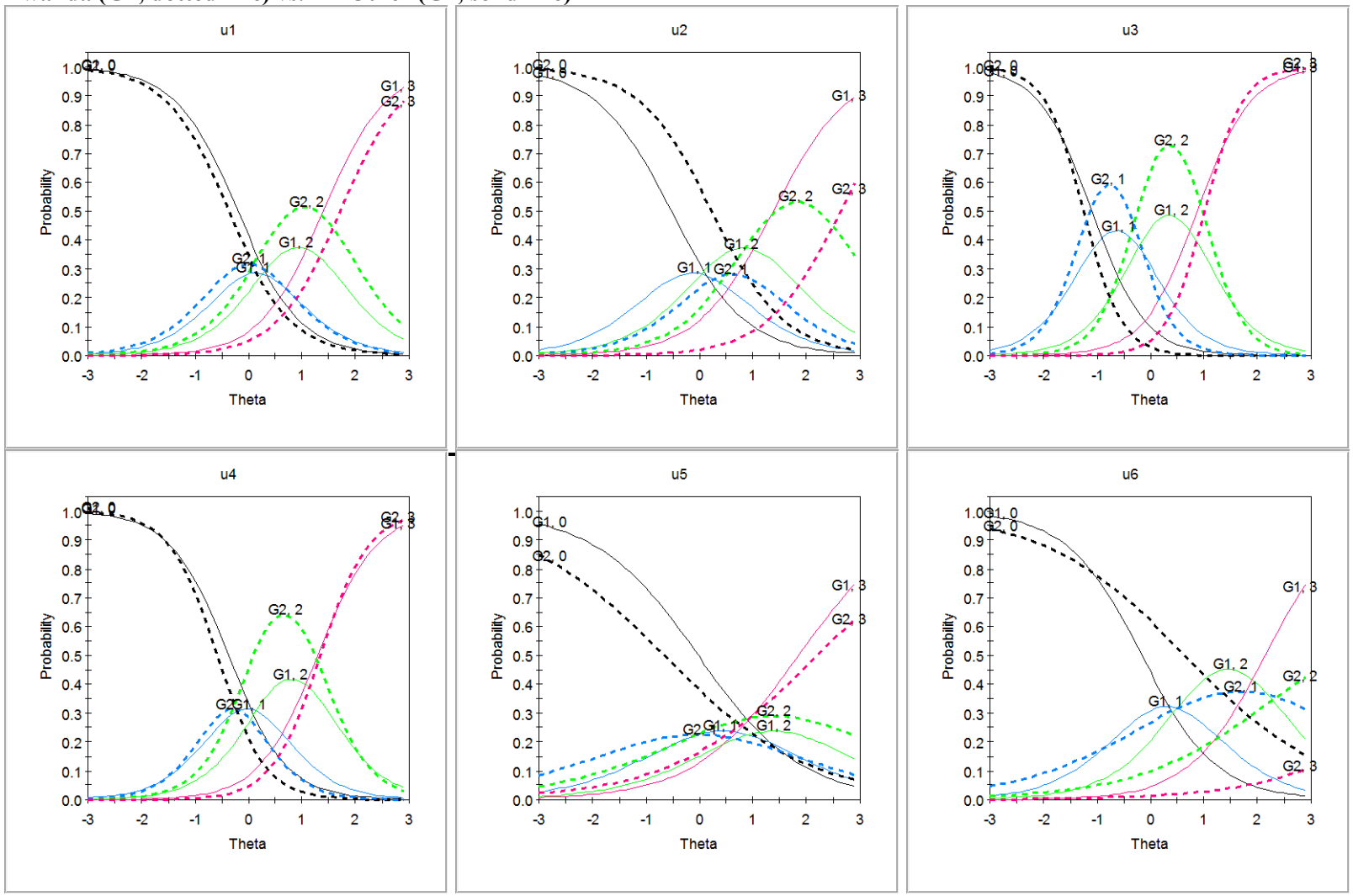

u13
u14
u15
Probability
Theta
G2, 0
G1, 0
G2, 1
G1, 1
G2, 2
G1, 2
G2, 3
G1, 3

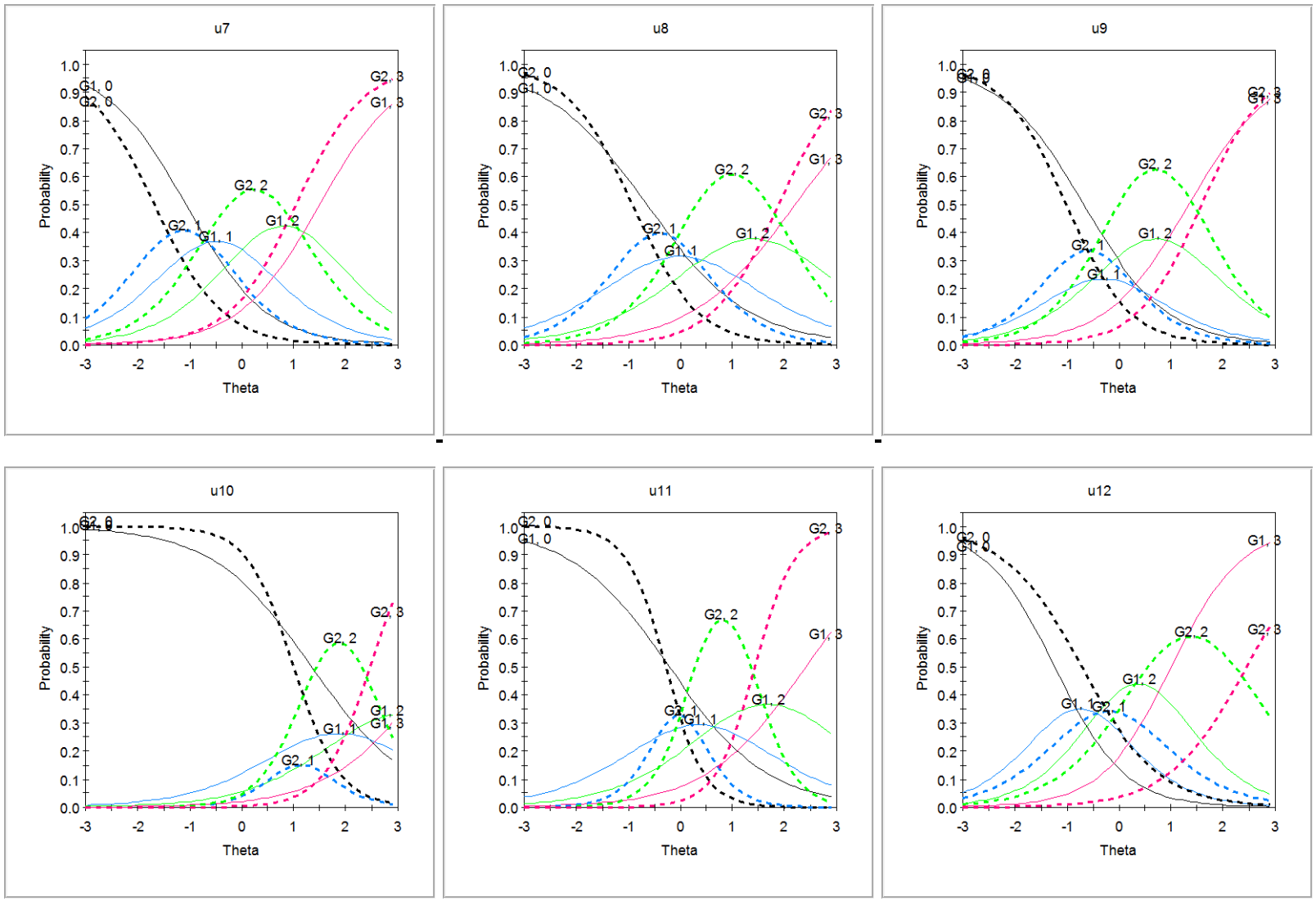
Figure A3.
Information curves for all items in Dohuk



Rwanda

Rwanda (G2, dotted line) vs. All Other (G1, solid line)




u7
u8
u9
u10
u11
u12
Probability
Theta
G1, 0
G2, 0
G2, 1
G1, 1
G2, 2
G1, 2
G2, 3
G1, 3

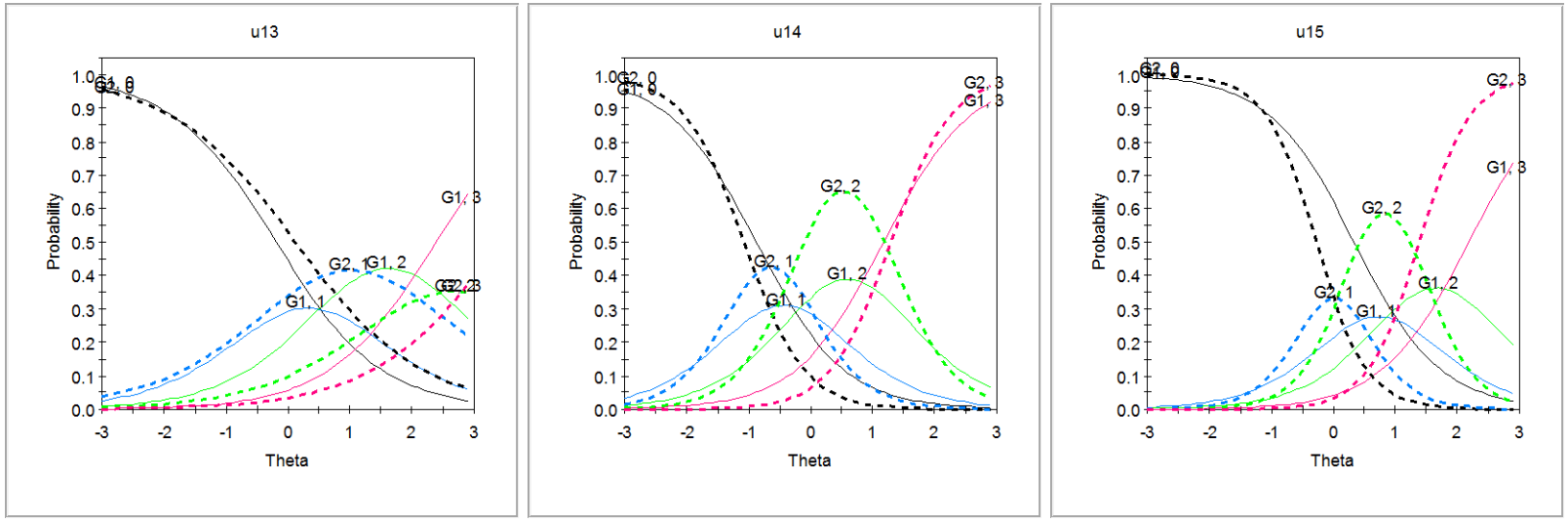
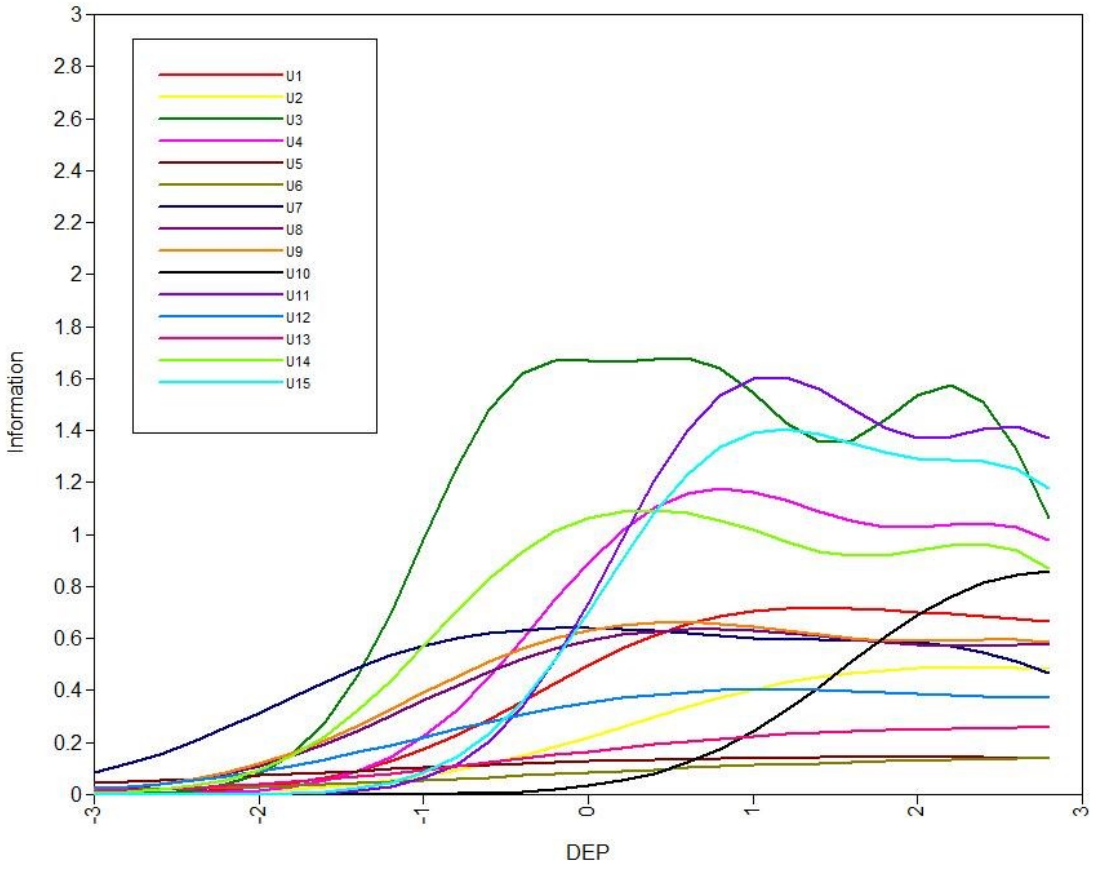
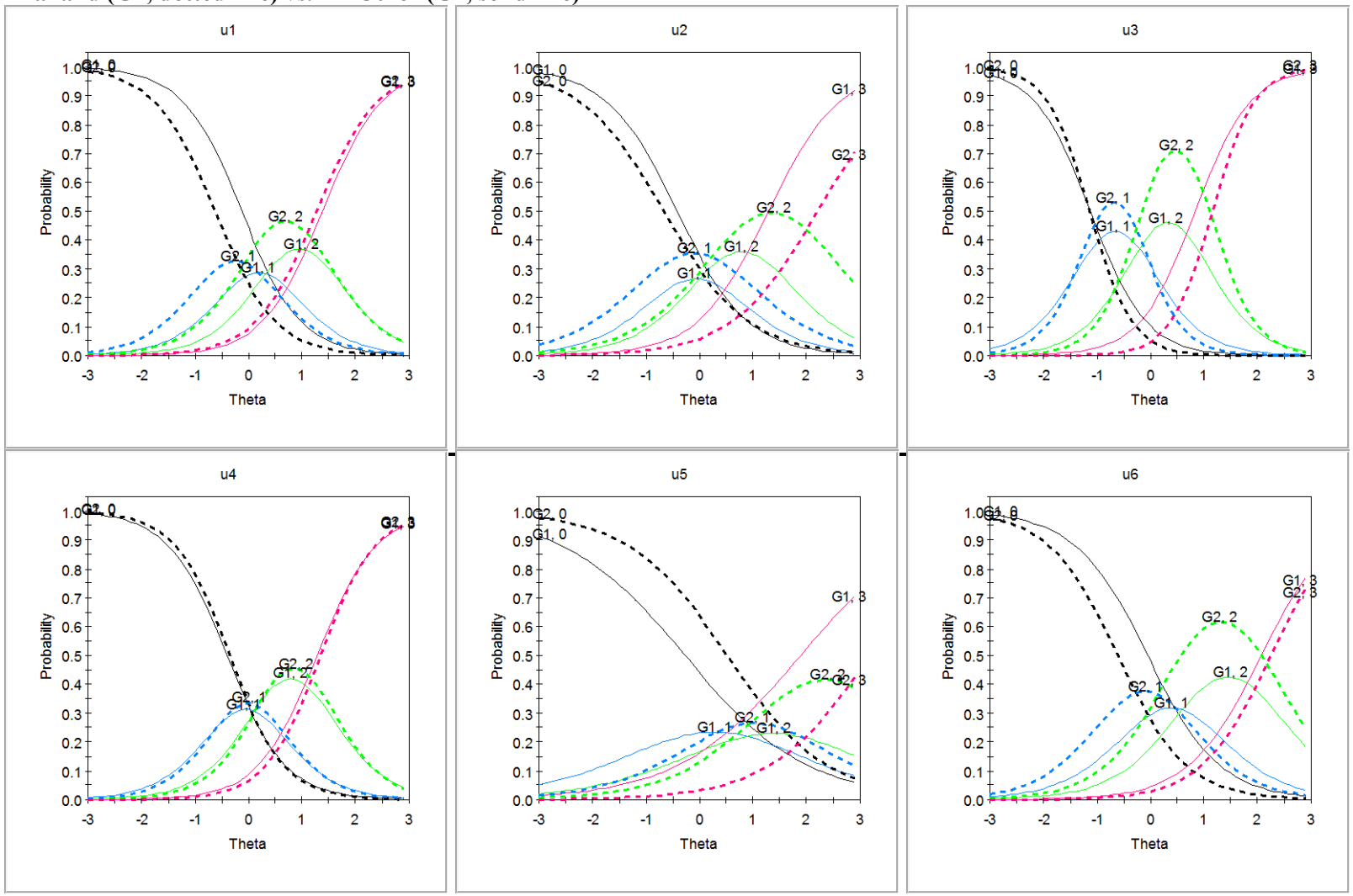

u13
u14
u15
Probability
Theta
G1, 1
G2, 1
G1, 2
G2, 2
G1, 3
G2, 3

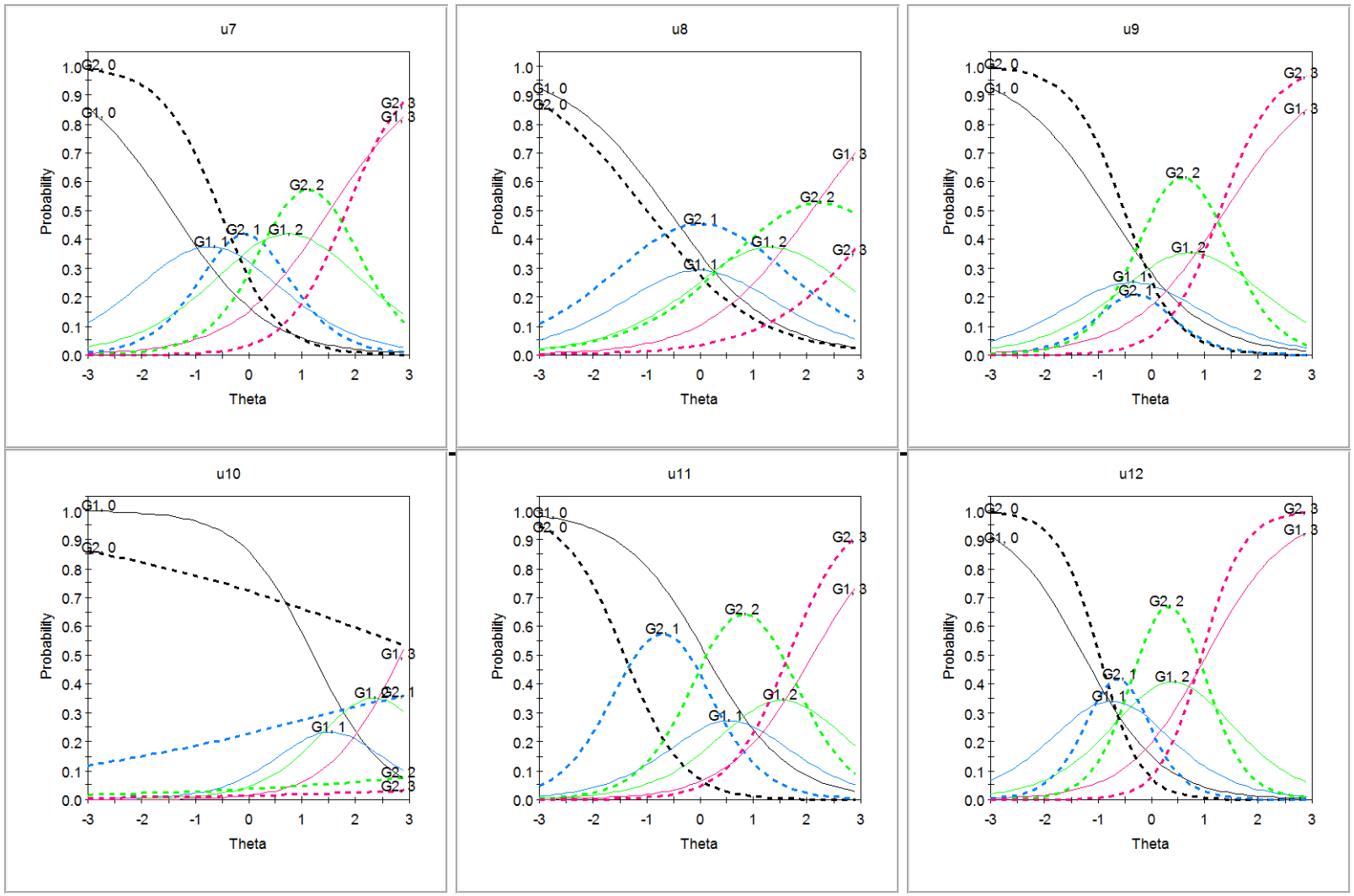
Figure A4.
Information curves for all items in Rwanda



Thailand

Thailand (G2, dotted line) vs. All Other (G1, solid line)




u7
u8
u9
u10
u11
u12
Probability
Theta
G1, 0
G2, 0
G1, 1
G2, 1
G1, 2
G2, 2
G1, 3
G2, 3

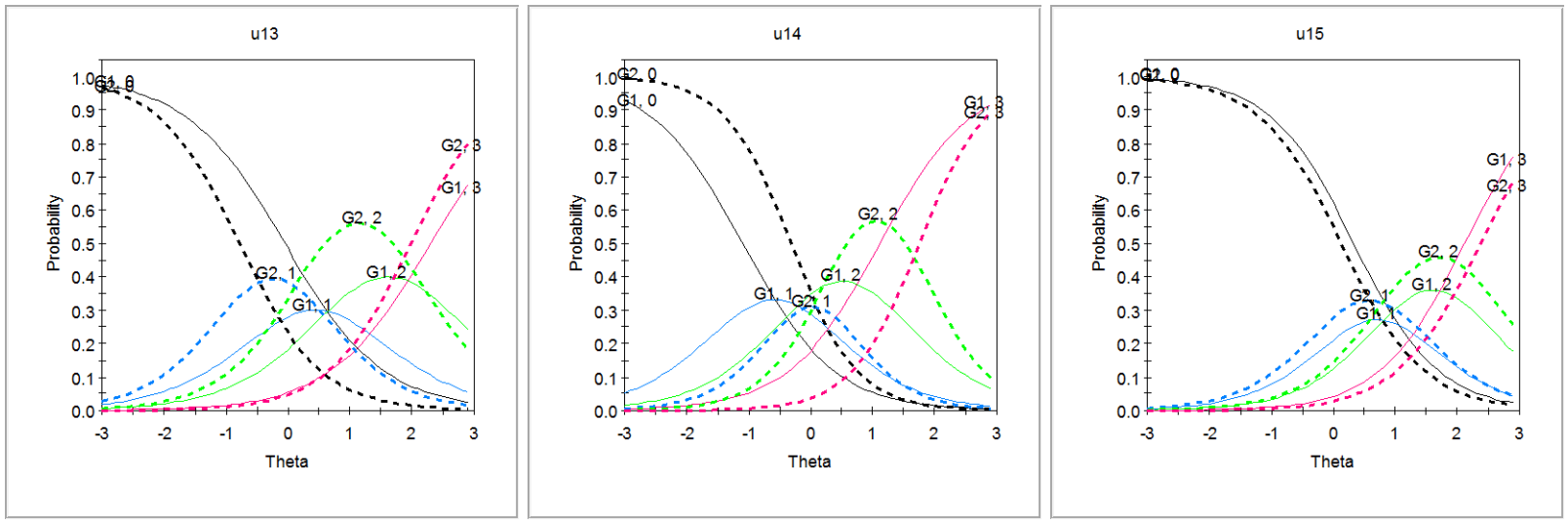
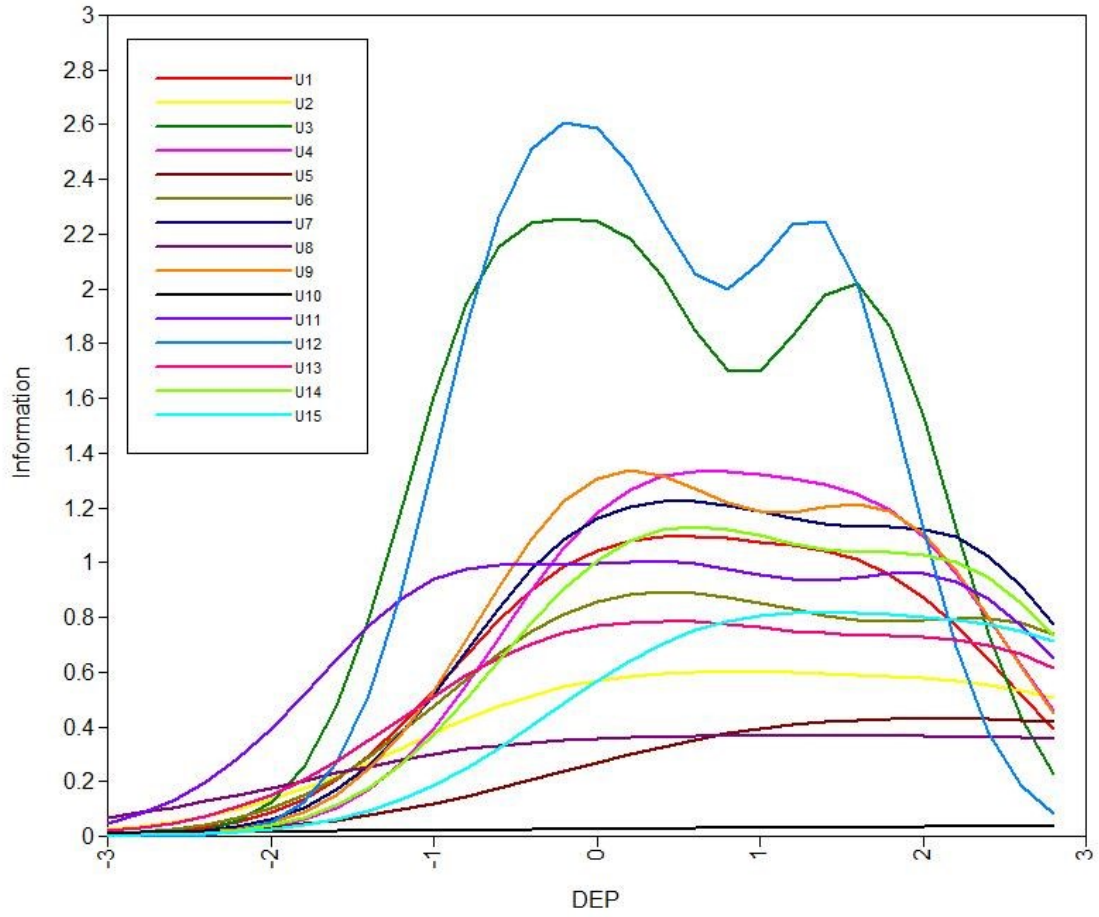
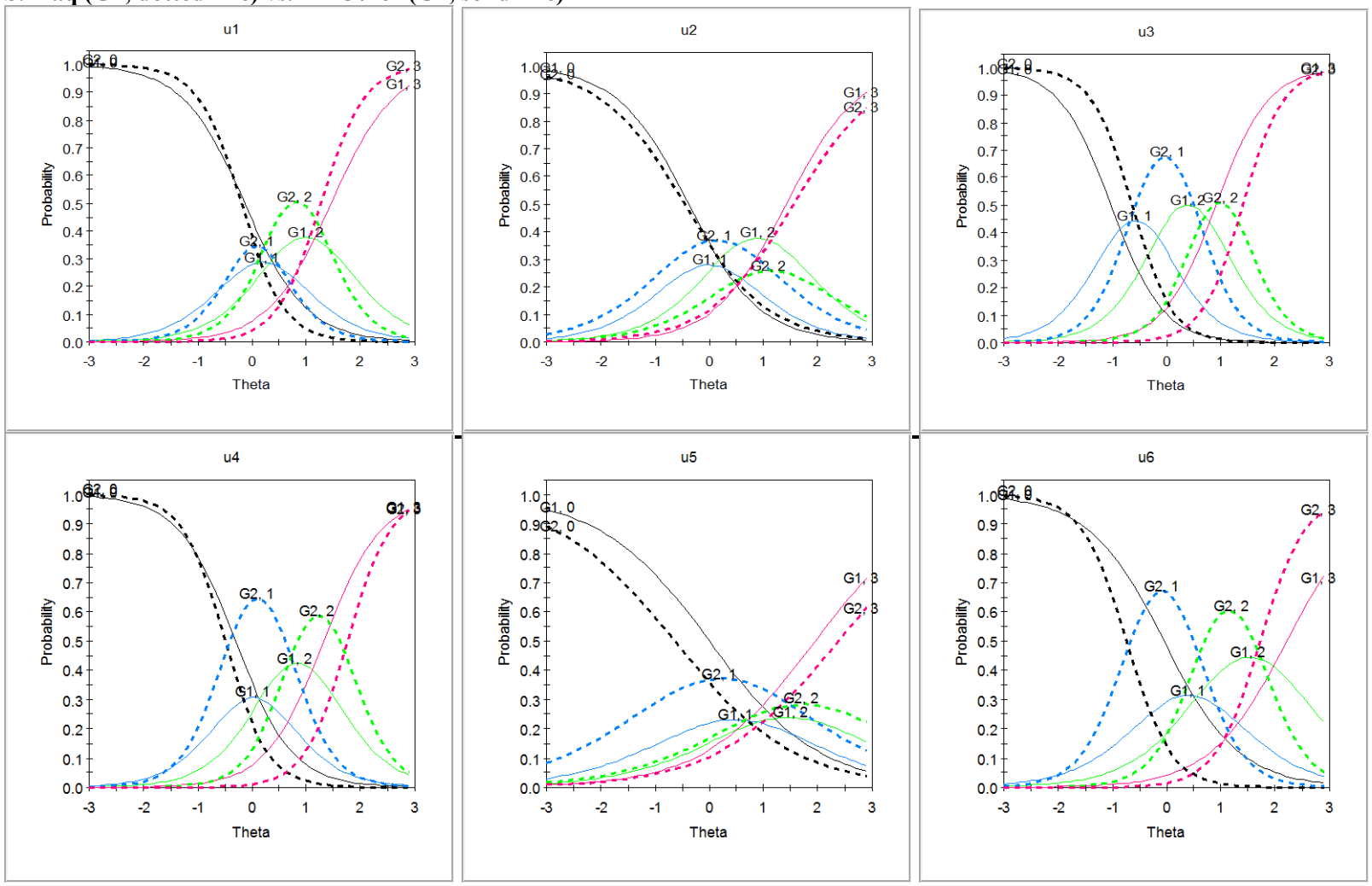

u13
u14
u15
Probability
Theta
G2, 0
G1, 0
G1, 1
G2, 1
G2, 2
G1, 2
G2, 3
G1, 3

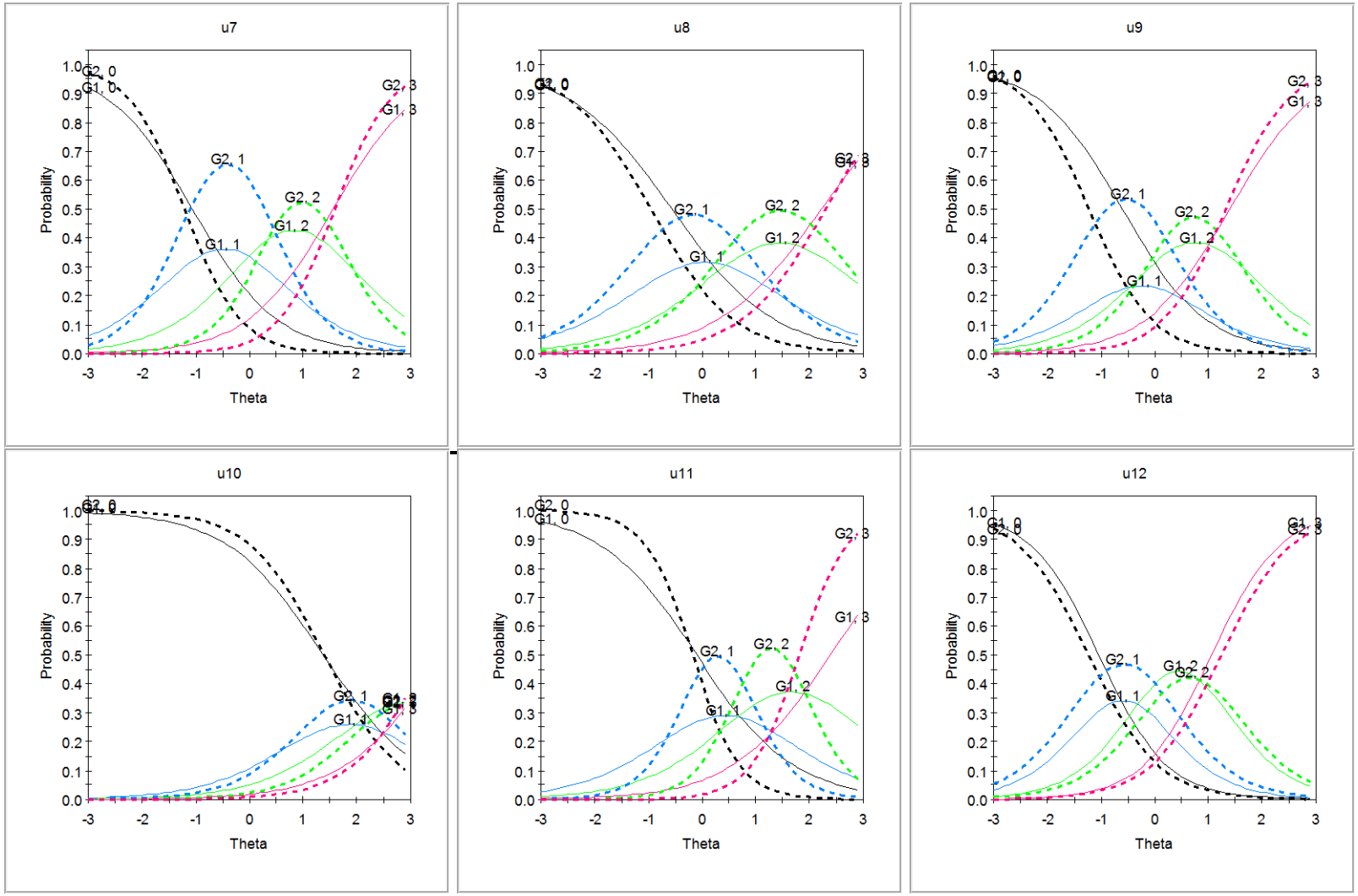
Figure A5.
Information curves for all items in Thailand



Southern Iraq

S. Iraq (G2, dotted line) vs. All Other (G1, solid line)




u7
u8
u9
u10
u11
u12
Probability
Theta
G2, 0
G1, 0
G2, 1
G1, 1
G2, 2
G1, 2
G2, 3
G1, 3

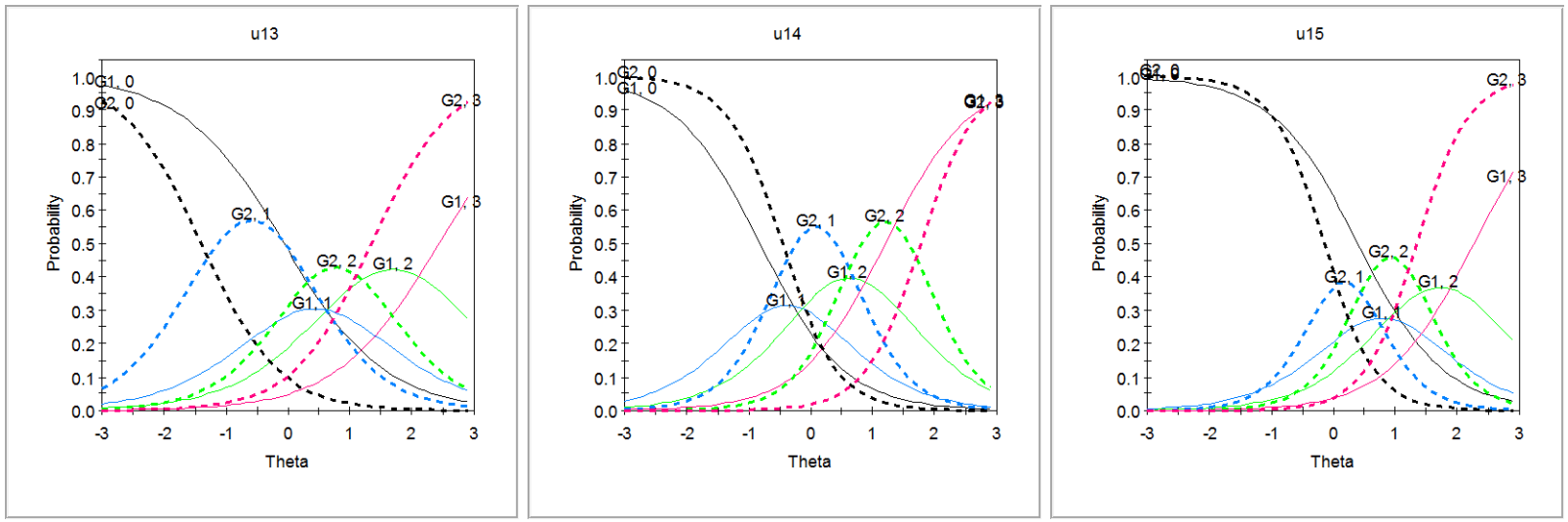
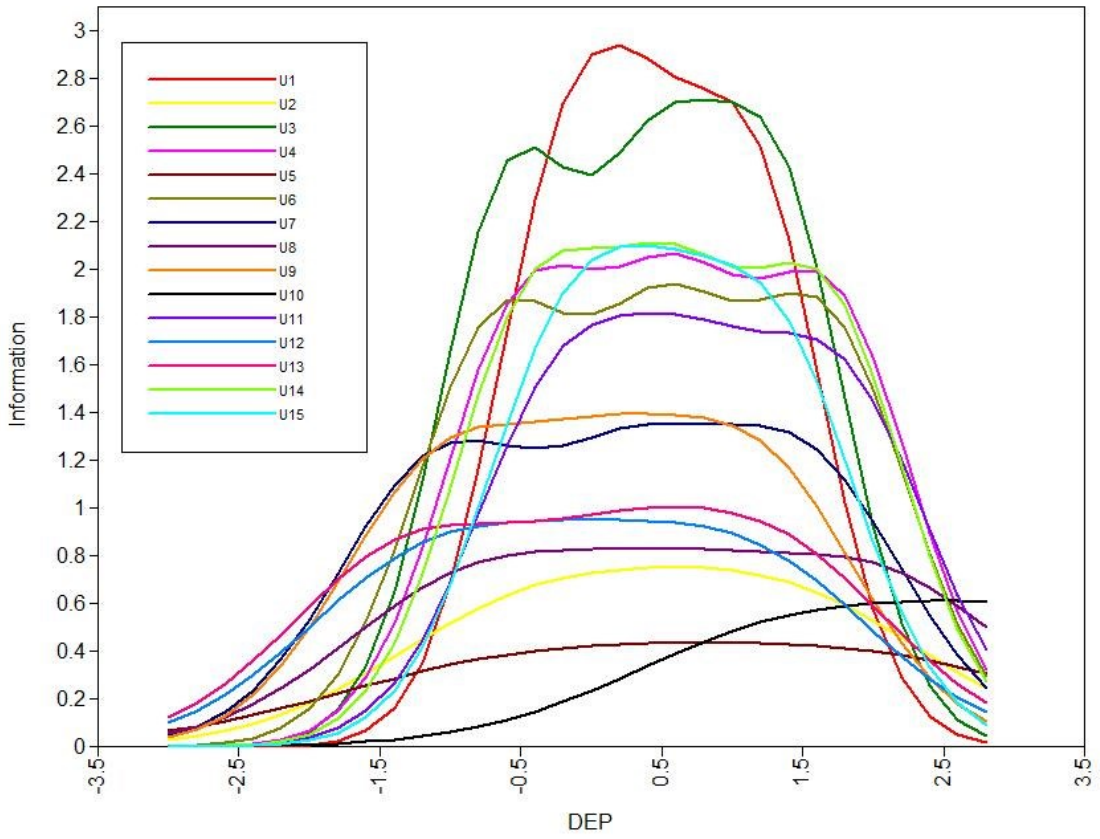
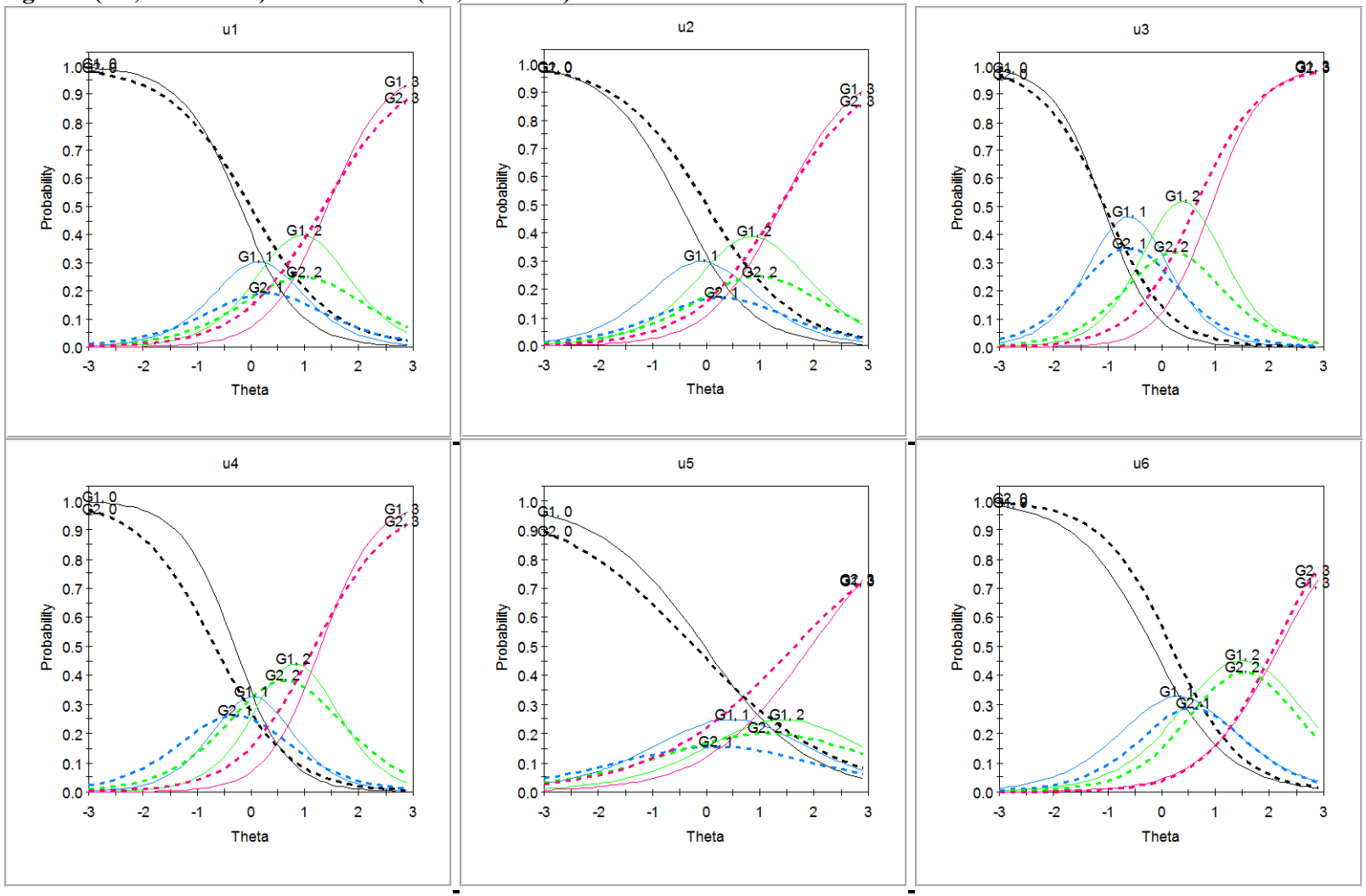

u13
u14
u15
Probability
Theta
G1, 0
G2, 0
G2, 1
G1, 1
G2, 2
G1, 2
G2, 3
G1, 3

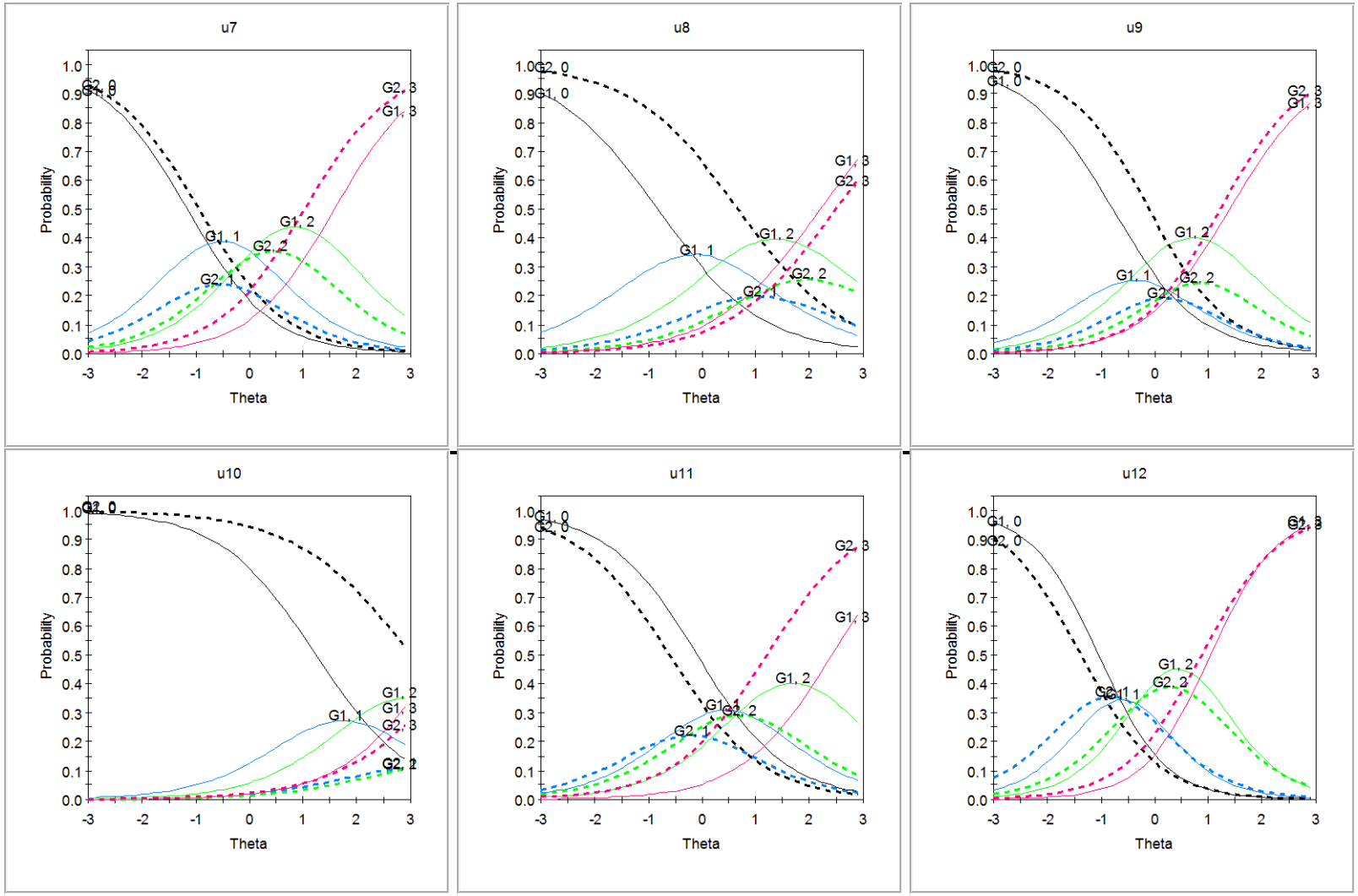
Figure A6.
Information curves for all items in S. Iraq



Uganda

Uganda (G2, dotted line) vs. All Other (G1, solid line)




u7
u8
u9
u10
u11
u12
Probability
Theta
G1, 0
G2, 0
G1, 1
G2, 1
G1, 2
G2, 2
G1, 3
G2, 3

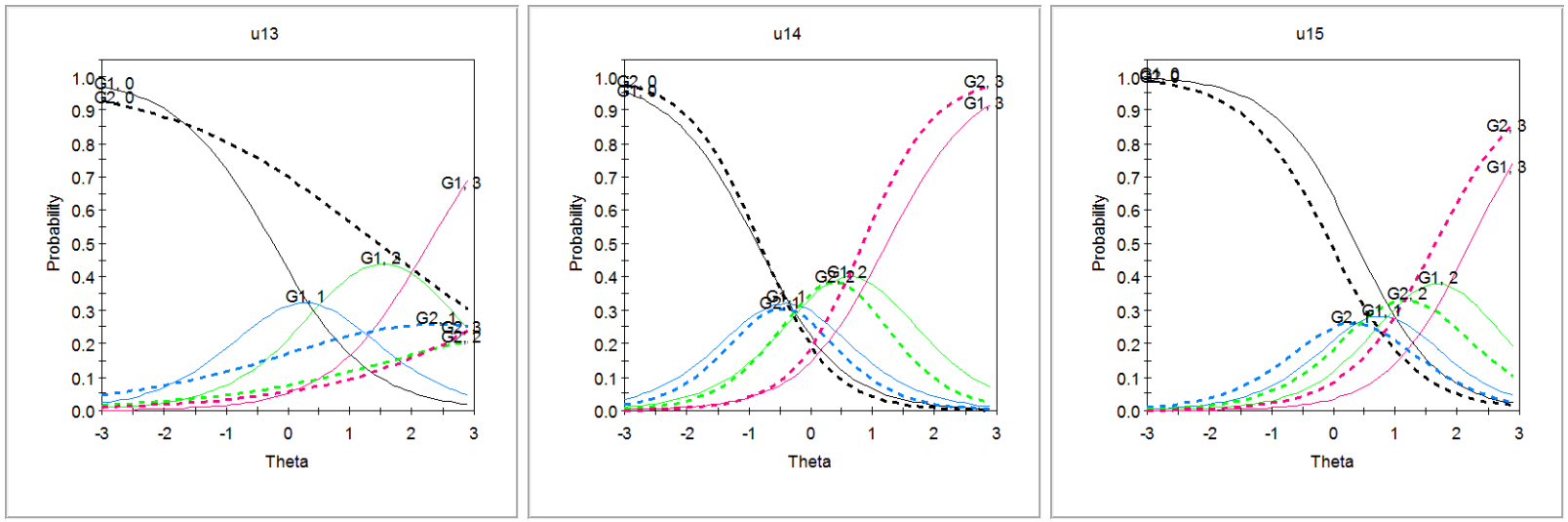
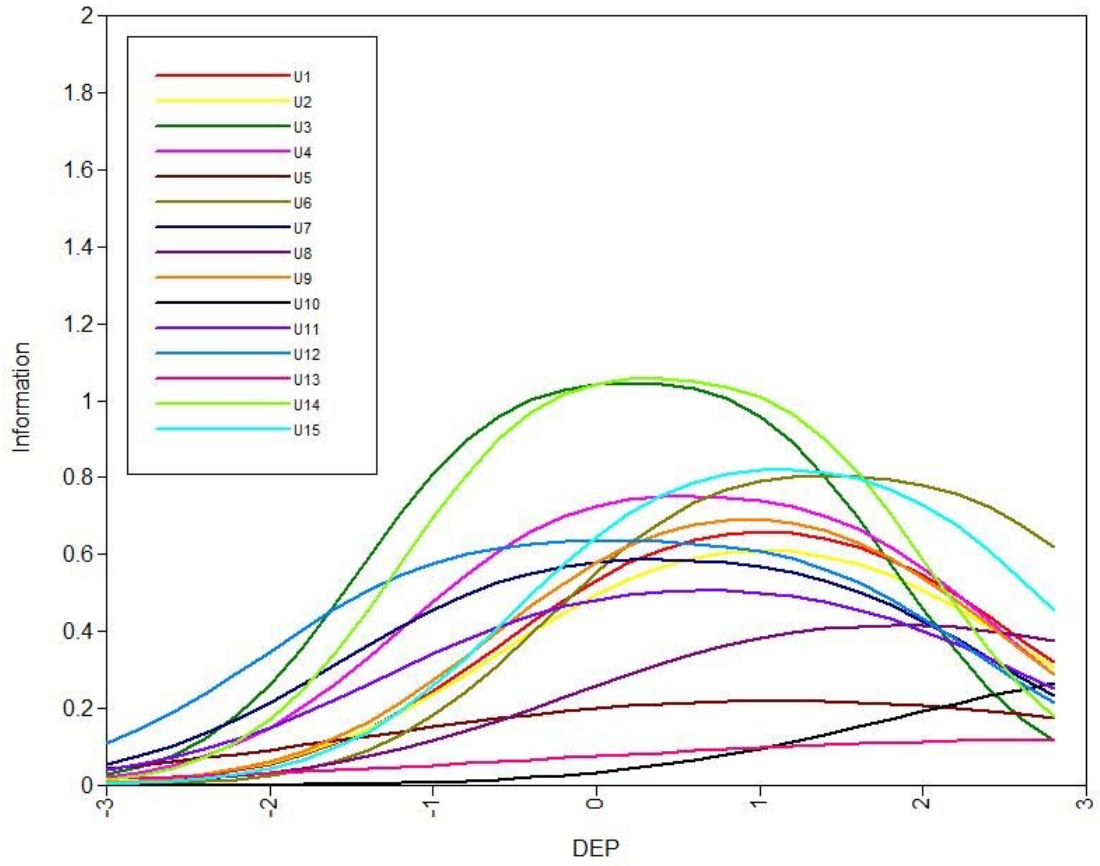
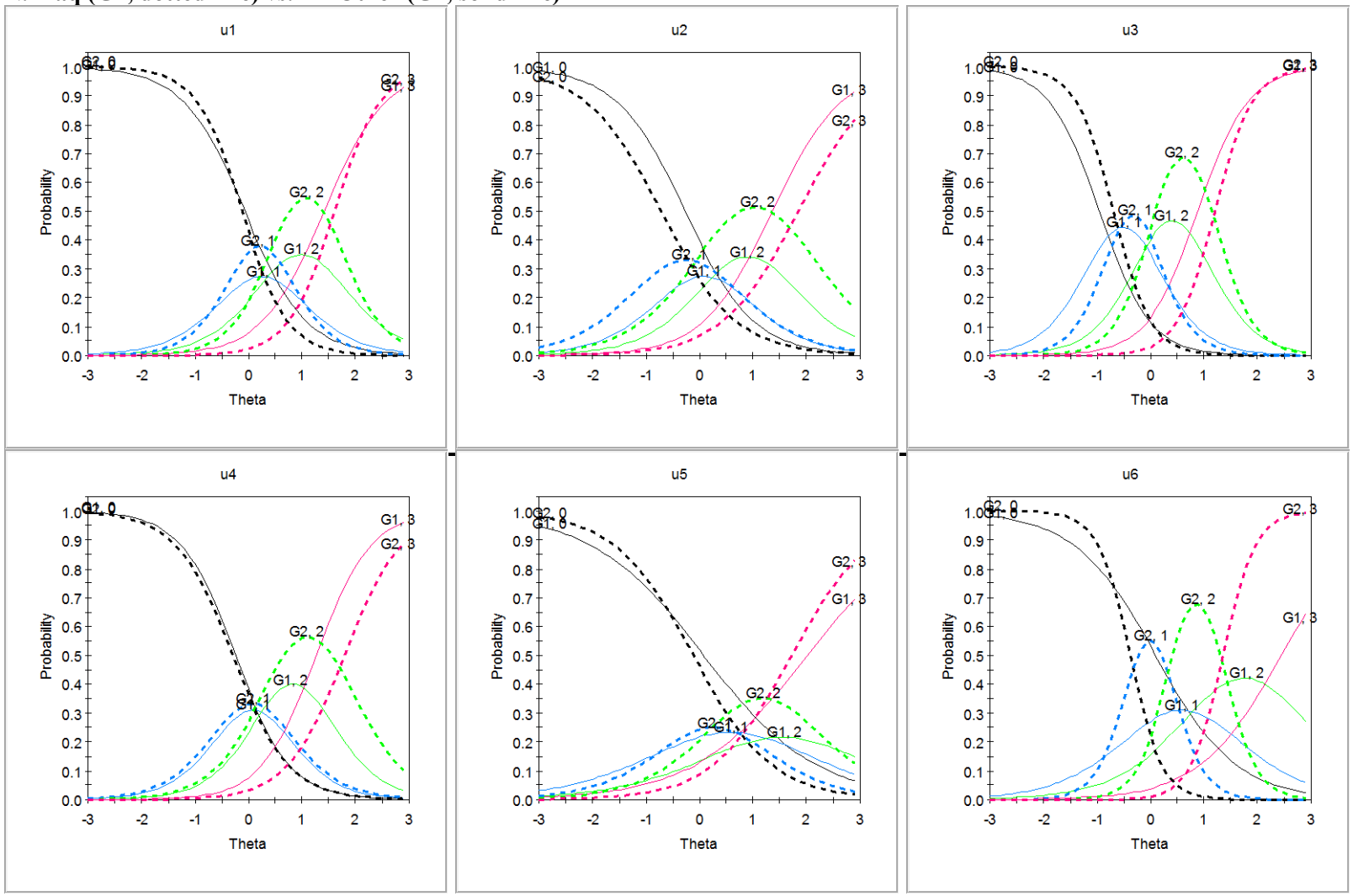

u13
u14
u15
Probability
Theta
G1, 0
G2, 0
G1, 3
G2, 3
G1, 2
G2, 2
G1, 1
G2, 1

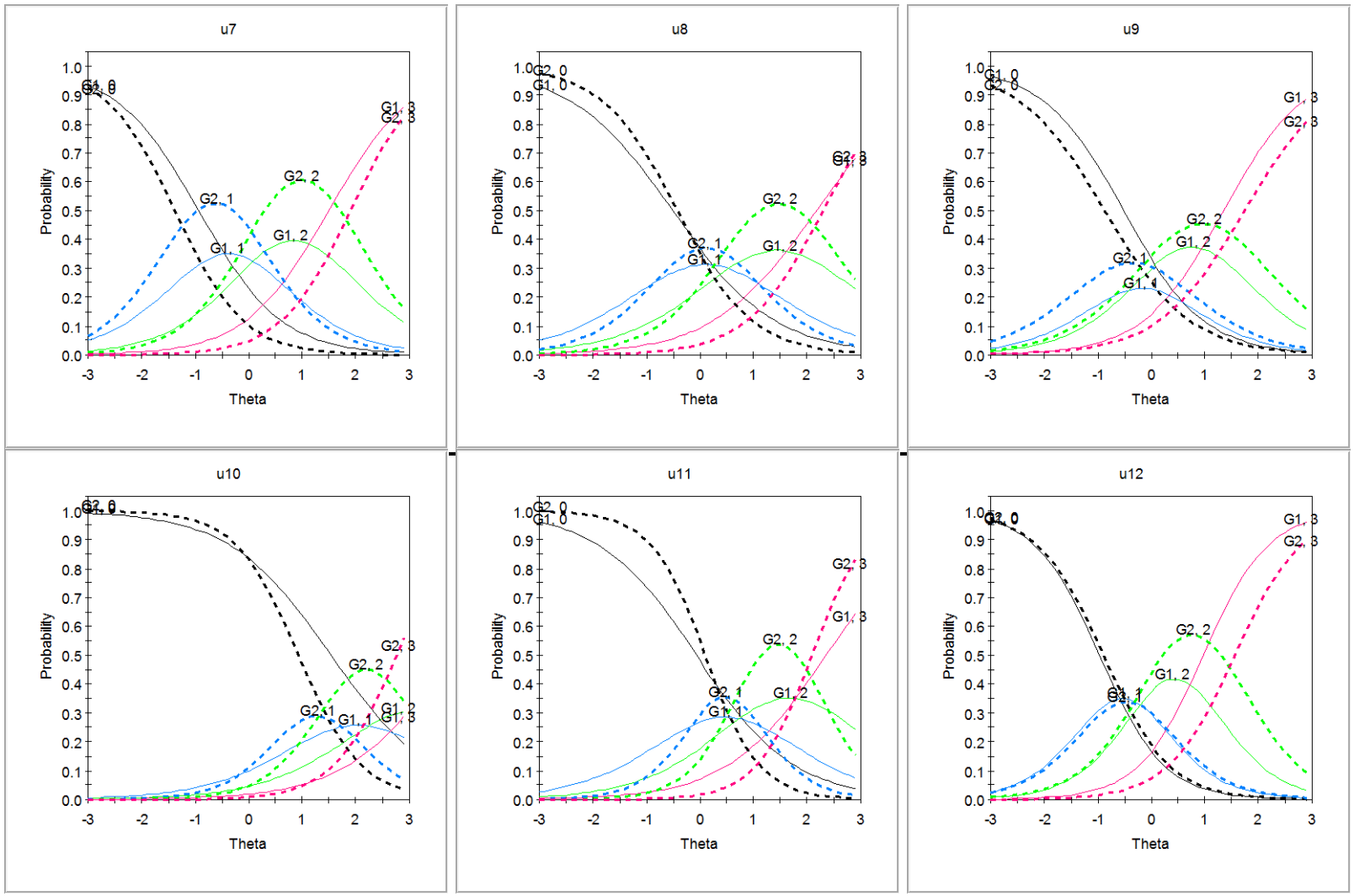
Figure A7.
Information curves for all items in Uganda



Northern Iraq

N. Iraq (G2, dotted line) vs. All Other (G1, solid line)




u7
u8
u9
u10
u11
u12
Probability
Theta
G1, 0
G2, 0
G1, 1
G2, 1
G1, 2
G2, 2
G1, 3
G2, 3

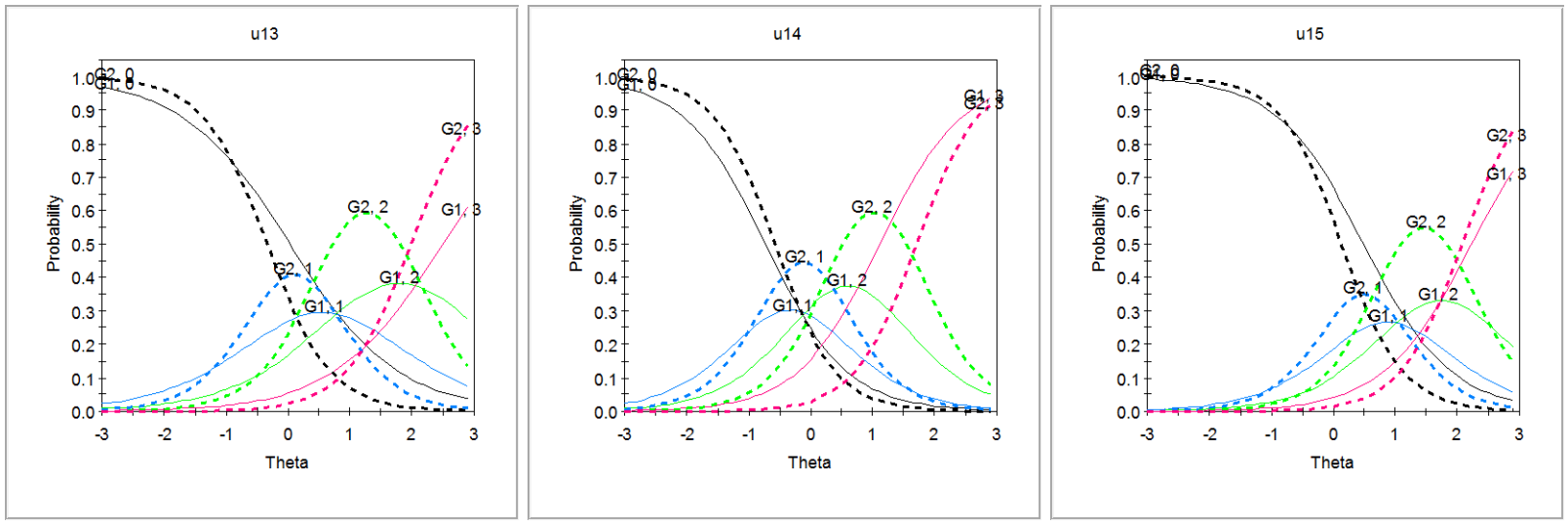
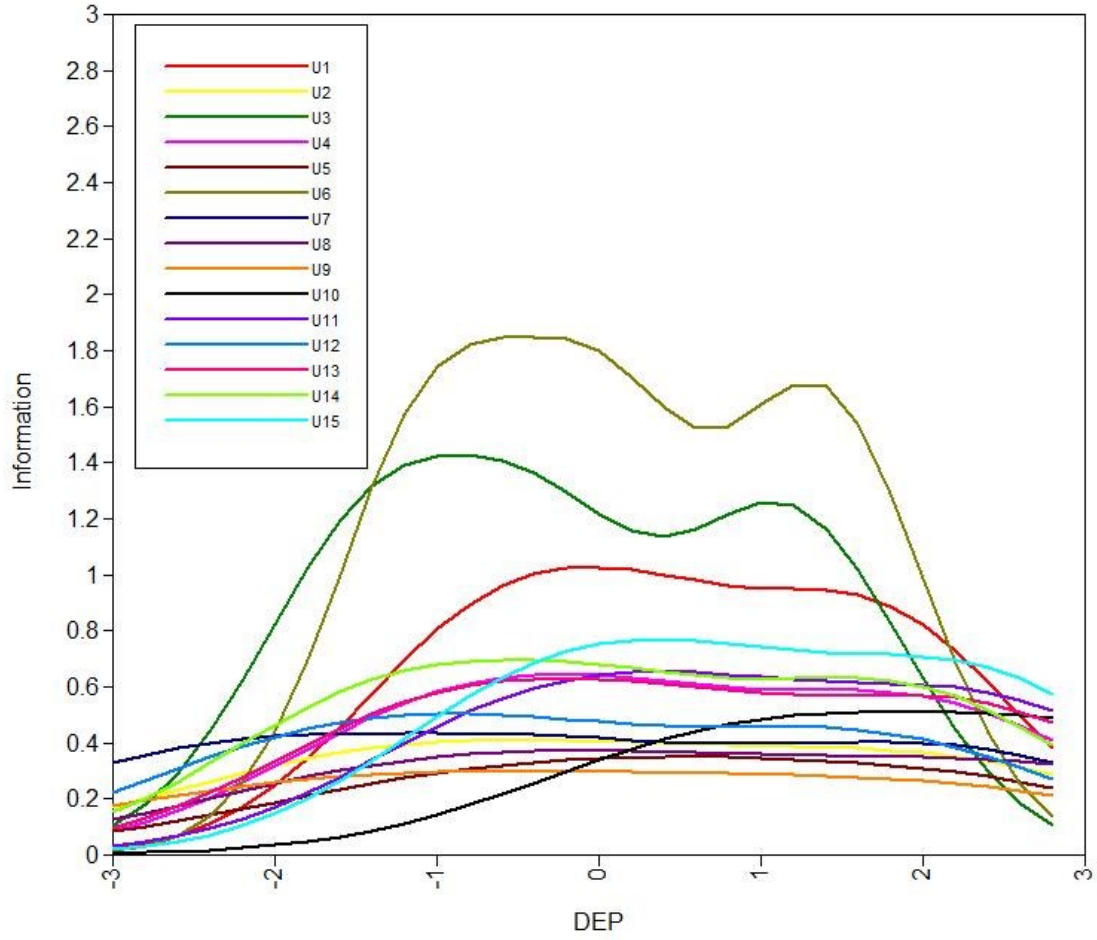

u13
u14
u15
Probability
Theta
G2, 0
G1, 0
G2, 1
G1, 1
G2, 2
G1, 2
G2, 3
G1, 3

Figure A8.
Information curves for all items in N. Iraq



Appendix E. Draft International Depression Symptom Scale – Global Version (IDSS-G)

IDSS-G

I would like to ask you questions about how things have been for you in the **last two weeks.** When you answer each question, I would like you to think back just over the **last two weeks**. In the past two weeks how often has each of the following problems occurred…

		None of the time	A little of the time	Most of the time	Almost all of the time
1	Feeling sad (H, Q, P)				
2	Feeling no interest in things/less interest in daily activities (H, Q, P)				
3	Crying easily (H, Q, P)				
4	Feeling hopeless about the future (H, Q, P)				
5	Feeling lonely (H, Q)				
6	Feeling socially withdrawn (Q, P)				
7	Feeling tired, low in energy or slowed down (H, Q, P)				
8	Weighing too little (Q, P)				
9	Weighing too much (Q)				
10	Problems with my appetite (H, Q, P)				
11	Problems with your sleep (H, Q, P)				
12	Feeling of being trapped or caught (H)				
13	Worrying too much about things (H, Q, P)				

14	Feelings of worthlessness (H, Q)				
15	Headache (Q)				
16	Stomach aches (Q)				
17	Other bodily aches and pains (Q, P)				
18	Feeling Angry (Q)				
19	Thinking too much (Q, P)				
20	Feeling confused (Q)				
21	Feeling weakness in your heart (Q)				
22	Heart palpitations (Q)				
23	Feeling as though your heart is heavy (Q)				
24	Feeling pressure on your heart (Q)				
25	Pain in your heart (Q)				
26	Moving or speaking so slowly or so fast that others have noticed (DSM)				
27	Difficulty concentrating (DSM, P)				
28	Difficulty doing your usual activities at home or work (Q)				
29	Thoughts of wanting to kill yourself (H, Q, P)				

*Items 28 and 29 will not be used in scoring
Q = from the qualitative review; H = from the HSCL; P = Also part of PTS measure; DSM = included based on DSM diagnostic criteria

Appendix F: Results from the IDSS - Local version

Descriptive statistics

Table A1.

Demographic information for instrument testing sample (N = 147)

Gender, *n* (%)	
Men	52 (35.4)
Women	95 (64.6)
Age, *M* (SD), *Range*	47.6 (13.6), *18-81*

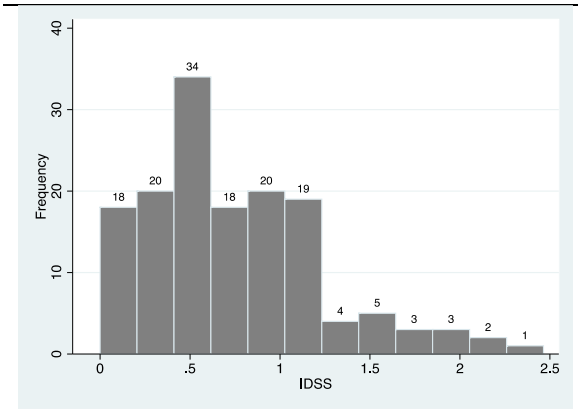
Table A2.

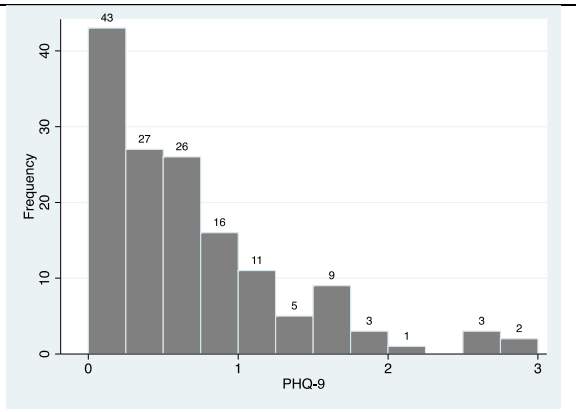
Mean scores and frequencies for each scale used in study

Measure	*N*	*M*	*Range*	*SD*	*Skew*
IDSS	147	0.73	0-2.46	0.50	1.03
PHQ-9	146	0.67	0-3	0.63	1.46
Functioning	147	0.61	0-2.43	0.60	1.08

Figure A1.

Histograms of summary scores on the IDSS, PHQ-9, and functioning scale





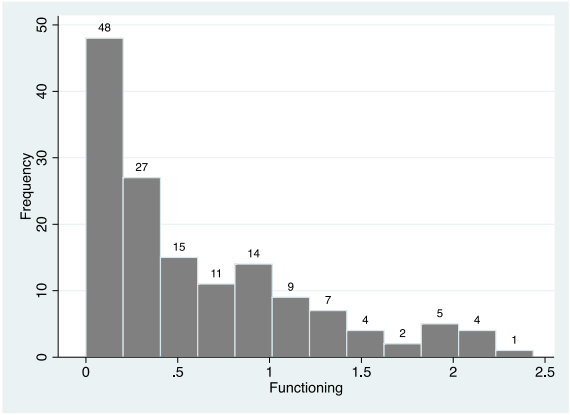


Table A3.

Frequency of diagnoses (N = 147)[a]

	N (%)
Any disorder	71 (48.3)
Depression	31 (21.1)
Dysthymia	39 (26.5)
GAD	22 (15.0)
None of these disorders	63 (42.9)
Co-morbidity (2 or more)	24 (16.3)

[a] Some individuals who were part of the analytic sample were diagnosed as having PTSD, but PTSD diagnoses were not included in the criterion validity analysis.

Reliability Results

Figure A2.

Scree plot with parallel analysis for items on the IDSS

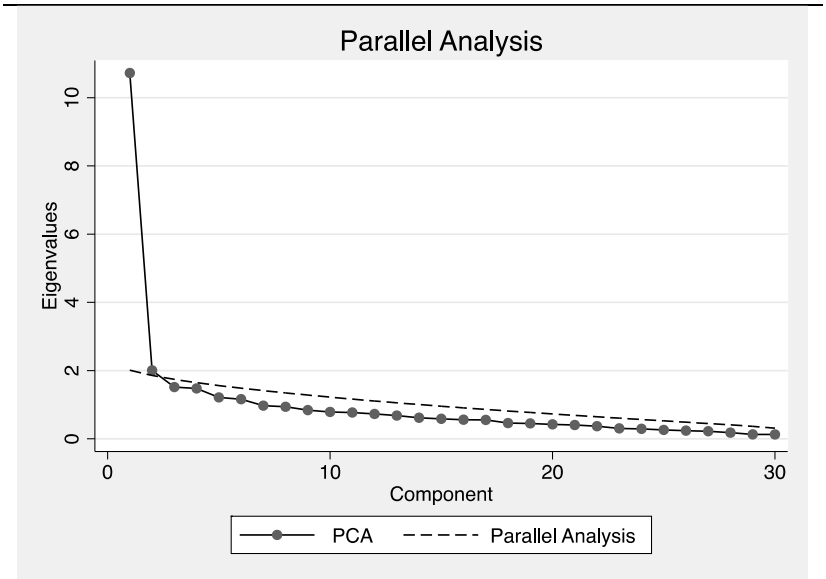


Table A4.

Model fit indices and their standard errors for various factor solutions for the IDSS

Model	X^2	Df	P value	RMSEA	CFI	TLI
1 factor model	665.58	405	0.000	0.057	0.909	0.903
2 factor model	536.39	376	0.000	0.043	0.944	0.936
3 factor model	461.01	348	0.002	0.035	0.961	0.951
4 factor model	397.52	321	0.010	0.025	0.973	0.964
5 factor model	317.96	270	0.024	0.019	0.979	0.970

Table A5.

Factor loadings for items on the IDSS

	F1	F2	F3
D01 Sad	0.712*	0.109	0.080
D02 no interest	0.676*	0.053	-0.068
D03 crying	0.559*	0.296*	0.054
D04 hopeless	0.567*	-0.043	0.229
D05 lonely	0.745*	0.029	-0.086
D06 social withdrawal	0.748*	0.130	-0.114
D07 tired/fatigue	0.260	0.336*	0.359*
D08 weigh too little	0.305	0.754*	-0.007
D09 weigh too much	0.058	-0.543*	0.131
D10 increased appetite	0.052	0.592*	0.288*
D11 sleep problems	0.270	0.218	0.288*
D12 trapped	0.898*	-0.019	-0.071
D13 worry	0.715*	-0.105	0.020
D14 worthless	0.574*	0.011	0.145
D15 headaches	-0.007	0.212*	0.579*
D16 stomach_aches	-0.284	0.332*	0.345*
D17 other_aches	0.207	0.153	0.230
D18 anger	0.592*	-0.280*	0.122
D19 thinking too much	0.809*	-0.233*	0.006
D20 confused	0.887*	-0.097	-0.068
D21 heart_weakness	0.059	0.159	0.540*
D22 palpitations	0.078	0.245	0.601*
D23 heavy_heart	0.003	-0.032	0.901*
D24 heart_pressure	0.091	-0.008	0.876*
D25 heart_pain	-0.043	0.340*	0.550*
D26 psychomotor	0.592*	0.247	-0.129
D27 concentration	0.618*	0.030	-0.086

D28 disappointed	0.714*	-0.039	0.162
D29 imp function	0.598*	0.089	0.130
D30 suicide	0.662*	0.329	0.053

Table A6.

Item analysis of depression items[a]

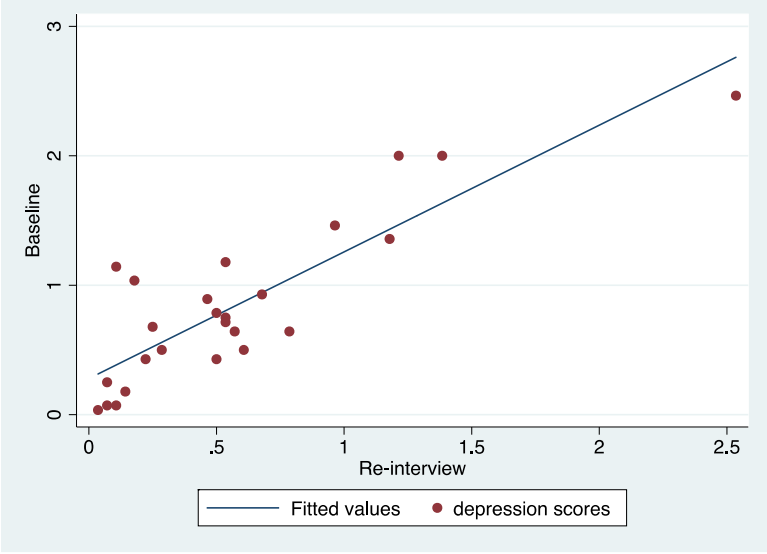
	# of obs	Sign	Item-test correlation	Item-rest correlation	Average inter-item covariance	alpha
D01 Sad	149	+	0.7369	0.7020	.2181959	0.9179
D02 no interest	146	+	0.5436	0.4962	.2268996	0.9211
D03 crying	148	+	0.6636	0.6303	.2252256	0.9195
D04 hopeless	147	+	0.6256	0.5854	.2245986	0.9199
D05 lonely	149	+	0.5847	0.5401	.2255933	0.9206
D06 social withdrawal	149	+	0.6129	0.5687	.2239172	0.9201
D07 tired/fatigue	149	+	0.6566	0.6152	.2220157	0.9194
D08 weigh too little	145	+	0.5621	0.5245	.2291946	0.9209
D09 weigh too much	144	-	0.0844	0.0511	.242401	0.9252
D10 increased appetite	149	+	0.5136	0.4663	.2285653	0.9216
D11 sleep problems	149	+	0.5820	0.5292	.2231358	0.9209
D12 trapped	149	+	0.7242	0.6891	.2194923	0.9182
D13 worry	149	+	0.6370	0.5920	.2220739	0.9198
D14 worthless	149	+	0.5675	0.5247	.2272442	0.9208
D15 headaches	149	+	0.5553	0.5018	.2247859	0.9214
D16 stomach_aches	149	+	0.1759	0.1253	.2405272	0.9257
D17 other_aches	149	+	0.4795	0.4214	.2280981	0.9227
D18 anger	149	+	0.4816	0.4365	.2306462	0.9220

D19 thinking too much	149	+	0.6302	0.5802	.2205224	0.9200
D20 confused	149	+	0.6853	0.6503	.222645	0.9189
D21 heart_weakness	149	+	0.5086	0.4569	.2278955	0.9219
D22 palpitations	148	+	0.6295	0.5912	.2250656	0.9198
D23 heavy_heart	148	+	0.6080	0.5711	.2268573	0.9202
D24 heart_pressure	148	+	0.6782	0.6480	.2255714	0.9193
D25 heart_pain	148	+	0.5014	0.4635	.2315424	0.9216
D26 psychomotor	147	+	0.5267	0.4835	.228871	0.9213
D27 concentration	148	+	0.4889	0.4339	.2278913	0.9222
D28 disappointed	148	+	0.7381	0.7054	.2191569	0.9179
Whole Scale					.2263812	0.9234

[a] The impaired functioning and suicidal ideation items were not included in the Cronbach's alpha analysis, since these items are not intended to be included in summary scores.

Figure A3.

Scatter plot of average IDSS scores at baseline and re-interview



Inter-Rater Reliability

For the IDSS-L, initial interviews and re-interviews with different interviewers were done with n = 30 individuals. On average re-interviews were done 10.2 days (SD = 5.3), with a range of 2-19 days, after the initial administration of the IDSS=L. The ICC across interviewers for average score on the IDSS was ICC = 0.90 with a 95%CI of [0.80, 0.95], indicating high inter-rater reliability.

Table A7.

Inter-rater reliability by pair of psychiatrist

Criterion	Pair 1 (n = 16)	Pair 2 (n = 26)
	Kappa (% agreement)	Kappa (% agreement)
MDD vs. all other diagnoses	0.86 (93.8)	0.78 (96.2)
Dysthymia vs. all other diagnoses	0.38 (68.8)	0.75 (92.3)
GAD vs. all other diagnoses	0.85 (93.8)	1.00 (100.0)
No diagnosis vs. all other diagnoses	0.64 (93.8)	0.91 (96.2)

Table A8.

Title and frequency of each of piles created during the pile sort activity[a]

Pile name	**N**
Feelings	16
Depression/sadness	13
Physical problems	11
Thinking too much	11

Related to disease	11
Heart Problems	10
Past/traumatic events	9
Stress	9
Trust/betrayal	9
Feeling angry	8
Dissatisfaction/disappointment	8
Functioning	6
Confusion	5
Feeling isolated/no one to rely on	5

[a] Piles that fewer than $n = 5$ people created were not included in the table

Table A9.

Frequency of each symptom on the depression and PTS instruments by pile

Symptom	Feelings (n = 16)	Depression (n =13)	Physical Problems (n =11)	Thinking too much (n =11)	Related to disease (n =10)	Heart Problems (n =10)	Past/traumatic events (n =9)	Stress (n =9)	Trust/betrayal (n =9)
D01 Sad	12	4	0	1	0	1	2	2	1
D02 no interest	9	1	0	1	4	0	4	1	0
D03 crying	8	2	3	2	2	0	1	1	3
D04 hopeless	9	4	1	0	0	1	1	2	2
D05 lonely	8	2	1	2	0	0	3	3	3
D06 social withdrawal	6	3	0	0	2	0	2	1	0
D07 tired/fatigue	6	3	6	0	2	0	2	1	0
D08 weigh too little	1	2	9	2	5	0	0	0	0
D09 weigh too much	1	3	8	1	4	0	2	1	0
D10 increased appetite	6	3	3	1	5	1	0	3	0
D11 sleep problems	7	4	5	1	1	1	3	1	2
D12 trapped	10	6	2	0	1	0	1	2	2

D13 worry	9	3	0	0	3	0	2	4	0
D14 worthless	9	2	1	0	0	0	3	2	1
D15 headaches	2	4	8	4	4	0	0	1	0
D16 stomach_aches	1	3	11	0	6	1	0	0	0
D17 other_aches	1	2	8	2	6	1	2	0	0
D18 anger	9	0	3	0	0	1	0	2	1
D19 thinking too much	11	2	0	7	2	0	2	3	0
D20 confused	9	5	1	2	2	0	1	2	1
D21 heart_weakness	4	2	5	1	4	5	1	0	0
D22 palpitations	2	5	8	0	6	6	0	0	0
D23 heavy_heart	1	5	8	0	5	5	0	0	0
D24 heart_pressure	1	4	7	0	6	6	0	0	0
D25 heart_pain	1	3	8	0	6	5	0	0	0
D26 psychomotor	4	0	3	3	0	1	4	1	0
D27 concentration	9	1	3	2	2	1	1	2	3
D28 disappointed	10	4	1	2	0	1	2	1	0
D29 imp function	3	1	3	0	1	0	2	0	0
D30 suicide	8	3	1	4	2	0	0	2	2

Validity results

Table A10.

Exploration of construct validity: Correlations of IDSS and other measured variables

	IDSS	Age	Gender	Functioning Scale	PHQ-9	Function Item	Suicide Item
IDSS	1.00						
Age	-0.16	1.00					
Gender	0.17	-0.06	1.00				
Functioning Scale	0.56*	-0.17*	-0.10	1.00			
PHQ-9	0.78*	-0.18*	0.06	0.50*	1.00		
Functioning Item	0.65*	-0.16	-0.05	0.48*	0.62*	1.00	
Suicide item	0.66*	-0.40*	0.09	0.50*	0.56*	0.56*	1.00

* $p < 0.05$

For item-level construct validity, all but four items significantly predicted more impairment in functioning after adjusting the *p*-value to account for multiple comparisons. The items that significantly predicted the most impaired functioning were "crying a lot" ($\beta = 0.33$) and "weighing too little" ($\beta = 0.31$). The effects of the other significant items ranged from $\beta = 0.15$ to $\beta = 0.28$. The items "weighing too much," "stomach aches," "other aches and pains" and "pain in the heart" did not significantly predict impaired functioning.

Table A12.

Average scale scores on depression measures by diagnostic category

	M (SD)	*Range*
Any disorder vs. no disorder		

Any disorder (n = 71)	*0.89* (0.48)*	*0.11-2.46*
No disorder (n = 63)	*0.55* (0.43)*	*0.00-2.11*
MDD/Dysthymia vs. no disorder		
Depression/Dysthymia (n = 52)	*0.95* (0.49)*	*0.11-2.46*
No disorder (n = 63)	*0.55* (0.43)*	*0.00-2.11*
GAD vs. no disorder		
GAD (n = 22)	*0.76* (0.40)	*0.29-2.00*
No disorder (n = 63)	*0.55* (0.43)	*0.00-2.11*
Depression vs. GAD		
Depressive disorders (n = 39)	*1.02* (0.52)	*0.11-2.46*
GAD (n = 18)	*0.73* (0.40)	*0.29-2.00*

* Paired t-tests indicate significant difference ($p < 0.05$) in means between groups

[a] Depression category includes both MDD and Dysthymia.

Table A13.

Effects of measured variables on impaired functioning presented as beta coefficients

Model	β (SE)	t
Model 1		
Age	-0.008 (0.01)	-2.21*
Model 2		
Age	-0.004 (0.01)	-1.22
Suicidal ideation[a]	0.71 (0.15)	4.75**
Model 3		
Age	-0.003 (0.01)	-0.92
Suicidal ideation	0.25 (0.16)	1.56

PHQ-9	0.37 (0.08)	4.78**
Model 4		
Age	-0.003 (0.01)	-0.92
Suicidal ideation	0.22 (0.16)	1.40
PHQ-9	0.13 (0.11)	1.21
IDSS	0.45 (0.14)	3.24**

[a] For the purposes of the incremental validity testing, the item related to suicide ideation was dichotomized meaning that 0 = none of the time and 1 = some, most and almost all of the time.

* $p < 0.05$

** $p < 0.001$

Table A14.

Area Under the Curves (AUC) for the IDSS and PHQ-9 across diagnostic categories

	IDSS	*PHQ-9*
	AUC, [95%CI]	*AUC,* [95%CI]
Any disorder vs. no disorder	0.72	0.74
	[0.63, 0.81]	[0.65, 0.82]
MDD/Dysthymia vs. no disorder	0.74	0.76
	[0.65, 0.83]	[0.67, 0.85]
GAD vs. no disorder	0.68	0.68
	[0.56, 0.79]	[0.56, 0.80]
Depression vs. GAD	0.30	0.33
	[0.15, 0.44]	[0.17, 0.50]

Table A15.

Cutoff values for average scores (range: 0-3) and corresponding classification statistics for the IDSS and PHQ-9

	IDSS Cutpoints			PHQ-9 Cutpoints		
	Optimal[a]	High (1.00)	Low (0.5)	Optimal (0.44)	High (1.11)	Low (0.3)
Any disorder vs. no disorder						
Sensitivity	0.70	0.35	0.80	0.82	0.27	0.89
Specificity	0.68	0.81	0.56	0.56	0.89	0.48
Positive Predictive Value	0.71	0.68	0.67	0.67	0.73	0.66
Negative Predictive Value	0.67	0.53	0.71	0.73	0.52	0.79
Correctly classified	0.69	0.57	0.69	0.69	0.56	0.69
MDD/Dysthymia vs. no disorder						
Sensitivity	0.75	0.44	0.83	0.89	0.31	0.92
Specificity	0.68	0.81	0.56	0.56	0.89	0.48
Positive Predictive Value	0.66	0.66	0.61	0.62	0.70	0.59
Negative Predictive Value	0.77	0.64	0.80	0.85	0.61	0.88
Correctly classified	0.71	0.64	0.68	0.74	0.63	0.68

[a] Optimal cutpoints for IDSS are 0.61 for Any vs. no disorder and 0.57 for MDD/Dysthymia vs. no disorder

Appendix G. Cognitive interview results: Most frequent meanings of each item

Table A1.
The most frequent meanings of each item asked about during cognitive interviews (*n*)

Item	1st most frequent	2nd most frequent	3rd most frequent
Feeling socially withdrawn	"When people do not want to talk to others. Even at home they are not very talkative with their family. Being socially withdrawn is having just few friends and not being interested in meeting new people" (12)	"Not wanting to interact with other people and don't want to talk to other people" (11)	"this is a feeling of depression or being unhappy…then I don't want to talk with anyone" (3)
Stomach aches	"This is a medical issue—having stomach pain because you are sick or have ulcers" (15)	"Stomach pain comes when you eat spicy food or feel unwell" (14)	"Stomach pain is a medical issue" (12)
Other bodily aches and pains	"Working a lot or being sick, then feel tired and body pain" (30)		n/a
Thinking too much	"This is when you have something in your mind or you are thinking about the current situation and politics and you cannot stop thinking about it. When your thoughts are too much" (26)	"thinking too much about their future" (7)	"It is a social issue. People think a lot about things they have to do in the present moment and also in their future and family's future" (3)
Feeling confused	"This is a social issue. When you have problems with friends and family and you don't know how to solve that problem you feel confused" (28)	"When unwanted or unexpected things happen to us and then we feel confused because we do not understand why those things happen" (6)	"When you have problems at work or lose your job, you feel confused" (5)
Feeling weakness in your heart	"This is a medical issue when your heart feels like it is not	"This can be related to chest pain and having a fast heartbeat" (15)	"I feel this way because I feel down" (4)

	strong enough" (27)		
Heart palpitations	"This is a medical issue" (28)	"If we feel stress, disappointment, or a lot of bad feelings then we can feel like this" (8)	"When people feel angry they can have stronger heat palpitations" (2)
Feeling as though your heart is heavy	"When we feel worried, our heart is feeling heavy" (17)	"This is just the symptom of things. This is some medical issue that is causing the problem" (14)	"I don't know how to describe this feeling because I never had it before" (7)
Feeling pressure on your heart	"This is a symptom of a medical problem" (26)	"When I feel stresses and anxiety I feel pressure on my heart" (12)	"Feeling your heart muscle is weak, being physically tired and having difficulty breathing" (7)
Pain in your heart	"This is a medical problem usually caused by high cholesterol" (30)	"It feels like your heart is being pierced by a needle" (11)	"When I feel pain in my heart I cannot breathe well" (2)
Moving or speaking so slowly or so fast that others have noticed	"When I am talking about something that I lie, my talking is fast" (25)	"This is a medical problem and the person is probably tired or overly excited" (12)	"when I am sad about something my speech can be slower than usual" (3)
Difficulty concentrating	"When you are busy and doing many things then it is difficult to concentrate" (13)	"When someone is not an expert on a specific field, it is difficulty for them to be concentrated" (11)	"This is a medical issue because they have difficulty looking at things longer and can't do that" (9)
Difficulty doing your usual activities at home or work	"This is just normal to have this problem if you don't have money or time to do things" (10)	"Having trouble completing tasks you usually do" (8)	"Sometimes when I feel a little bit down, depressed and I do not have enough energy I have some difficulty at doing my regular activities, I feel like I do not want to do anything neither at home nor at work" (7)

CURRICULUM VITAE

EMILY E. HAROZ, M.A., M.H.S.

D.O.B. December 30, 1981 in Boston, MA

Personal Information

Permanent Address:
407 Bretton Pl
Baltimore MD, 21218

University Address:
Johns Hopkins Bloomberg School of Public Health
Department of Mental Health
624 N. Broadway
Baltimore MD, 21205

Email: eharoz1@jhu.edu
Phone: 617-894-8168
Fax: n/a
Skype : emilyharoz
Citizenship: United States

Current Positions/Training

2011-present **Psychiatric Epidemiology Training Program Fellow**
National Institute of Mental Health and Johns Hopkins School of Public Health
Training provided in combination with the departments of Epidemiology and Biostatistics, and the department of Psychiatry and Behavioral Sciences at the school of medicine.
Stipend and full tuition coverage for four years

Education

2011-present **Johns Hopkins School of Public Health,** Baltimore MD
Doctoral Candidate, Department of Mental Health

2010-2011 **Johns Hopkins School of Public Health,** Baltimore MD
M.H.S. Department of Mental Health (May 2011)

2008-2009 **Colombia University,** New York, NY
M.A. Psychology in Education (December 2009)

2000-2004 **University of Puget Sound,** Tacoma, WA
B.A. History, (May 2004)

Honors & Awards

2005-2007 **Pearson Teaching Fellow**
Fellowship in teaching from Pearson, Inc. and Jumpstart for Young Children to teach in an early childhood center that serves children in low-income communities.

2003-2004 **Order of Omega**
Leadership honor society for juniors and seniors who have exemplified high standards in the area of leadership and involvement within their respective organizations and local community.

2000-2001 **SPURS Honor Society**
Honor society for freshman to promote academic excellence and service to the campus and community.

Peer-Reviewed Journal Articles

1. **Haroz EE**, Murray LK, Bolton P, Betancourt T, Bass JK. Adolescent resilience in northern uganda: The role of social support and prosocial behavior in reducing mental health problems. *J Res Adolesc*. 2013;23(1):138-148.
2. Bass JK, Ayash C, Betancourt TS, **Haroz EE**, et al. Mental health problems of displaced war-affected adolescents in northern uganda: Patterns of agreement between self and caregiver assessment. *J Child Fam Stud*. 2013:1-11.
3. Murray LK, Dorsey S, **Haroz EE**, et al. A common elements treatment approach for adult mental health problems in low-and middle-income countries. *Cognitive and Behavioral Practice*. 2014;21(2):111-123.
4. **Haroz EE**, Ybarra ML, Eaton WW. Psychometric evaluation of a self-report scale to measure adolescent depression: The CESDR-10 in two national adolescent samples in the united states. *J Affect Disord*. 2014;158:154-160.
5. Kohrt BA, Rasmussen A, Kaiser BN, **Haroz EE**, et al. Cultural concepts of distress and psychiatric disorders: Literature review and research recommendations for global mental health epidemiology. *Int J Epidemiol*. 2014;43(2):365-406.
6. **Haroz EE**, Bass JK, Lee C, Murray LK, Robinson C, Bolton P. Adaptation and testing of psychosocial assessment instruments for cross-cultural use: An example from the thailand burma border. *BMC Psychology*. 2014;2(1):31.
7. Bolton, P. Lee, C, **Haroz, EE,** Murray, L. Dorsey, S., Ugueto, A., Robinson, C. Bass, J. A Transdiagnostic Community-Based Mental Health Treatment for Comorbid Disorders: Development and Outcomes of a Randomized Controlled Trial among Burmese Refugees in Thailand. *PLOS Medicine.* 2014.
8. Michaleopoulous, L., Unick, J., **Haroz, E.E.**, Bass, J. Murray, L., & Bolton, P. An exploration of the construct validity of PTSD based on data from three highly diverse populations. *Traumataulogy*. 2014.
9. **Haroz EE**, Jordans MJD, de Jong JTVM, Gross A., Bass JK, Tol WA. Measuring hope, a construct of resilience among children affected by armed conflict: Cross-cultural construct validity of the Children's Hope Scale. (in revision)
10. Kaiser B, **Haroz EE,** Kohrt B, Bass J, Bolton P, Hinton D. "Thinking too much:" A systematic review of a common idiom of distress. (in revision).
11. Nguyen, A., **Haroz, E.E.**, & Bass, J. *Correlates and symptom profiles of maternal mental distress in three low-income countries*. (Under review).
12. Lewandowski, E., Verdeli, H., Feighery, A. Bass, J., Hamba, C. **Haroz, E.E.**, Stavrou, V., Ndogoni, L. & Bolton, P. (in preperation). Local Perceptions of Group Interpersonal Psychotherapy in Rural Uganda. *Social Science and Medicine*.

Book Chapters

1. Tol WA, **Haroz EE**, Hock RS, Kane JC, Jordans MJ. Ecological perspectives on trauma and resilience in children affected by armed conflict. *Helping Children Cope with Trauma: Individual, Family and Community Perspectives*. 2014:193.

Nongovernmental Organization (NGO) Reports

1. **Haroz EE**, Lee, C, Bolton, P, Robinson C. 2011. *Assessment of Survivors of Torture and Violence from Burma living in Thailand: Development and Testing of a Locally-Adapted Psychosocial Assessment Instrument*, USAID, Washington, D.C.
2. Lee C, **Haroz EE**, Bolton P, Bass J, Robinson C, Murray L, Dorsey S. 2012. *Assessment of the impact of a common elements treatment approach for survivors of torture and violence from Burma living in Thailand*, USAID, Washington, D.C.

Other Publications

1. **Haroz EE**. 2012. Development and testing of a locally-adapted psychosocial assessment instrument. *International Society for Traumatic Stress Studies: Traumatic Stress* Points. March 16(2):10-12.

Presentations/Abstracts

1. Augello, J., Soloman, Beylul, **Haroz, E.**, Hajcak, G., Bonnano, G. & Dennis, T. (2009, May) *The late positive potential is sensitive to individual differences in trait anxiety.* Poster presented at the Association of Psychological Science annual convention. San Francisco, CA.
2. Hong, M.**,** Augello, J., **Haroz, E.** & Dennis, T. A. (2009, October). *The late positive potential as a biomarker for gender differences in affective processing related to emotion regulation.* Poster presented at the Society for Psychophysiological Research Annual Convention. Berlin, Germany.
3. Powers, A., Saunders, L., Kessel, E., Dobrich, O., **Haroz, E.** & Dennis, T. A. (2010, May). *The late positive potential varies with observed behavioral inhibition in school-aged children*. Poster presented at the Association for Psychological Science annual convention. Boston, MA.
4. Scott, C., Dennis, T., DeCicco, J., Saunders, L., Kessel, E., Dobrich, O., Powers, A., & **Haroz, E.** (2010, November). *Neural and temperamental correlates of fearful behavior in children*. Poster presented at the Annual Biomedical Research Conference for Minority Students. Charlotte, NC.
5. **Haroz, E.** (April, 2012). *Development of a Psychosocial Instrument for Burmese Survivors of Systematic Violence Displaced in Thailand*. Oral presentation at the annual meeting for the American Pscyhopathology Association. New York, NY.
6. **Haroz, E.** Bass, J., Lee, C. Murray, L. K., Robinson, C. & Bolton, P. (2012, November). *Development and testing of an adapted psychosocial assessment instrument with survivors of systematic violence from Burma*. Panel presentation at the annual conference for the International Society of Traumatic Stress Studies. Los Angeles, CA.
7. **Haroz, E.**, Weiss, W., Mahmooth, Z., & Bolton, P. (2012, November). *A common elements transdiagnostic approach (CETA) for trauma affected populations in Southern Iraq: Preliminary data from a randomized controlled trial.* Panel presentation at the annual conference for the International Society of Traumatic Stress Studies. Los Angeles, CA.
8. **Haroz, E.** Bass, J, & Bolton P. (2014, March). *Depression symptoms across contexts: Development of a cross-contextually valid measure of depression*. Panel presentation at the annual conference for the Society of Applied Anthropology. Albuquerque, NM.
9. **Haroz, E.** & Michalopoulos, L. (2014, July). *Commonalities in cross-cultural symptomatology for depression and posttraumatic stress disorder*. United States Agency for International Development Victims of Torture Meeting. Washington, DC.
10. **Haroz, E.**, Nguyen, A., Murray, L, Bass, J., & Bolton, P. (submitted). *Development and testing of implementation measurement instruments: A mixed methods study in Iraqi Kurdistan and Myanmar*. Annual meeting of the Society for the Study of Psychiatry and Culture. Providence, RI.

Referee & Editorial Experience

Reviewer	*Conflict and Health*
	Global Public Health
	Journal of Nervous and Mental Disease (JNMD)
	Transcultural psychiatry

Professional Experience

Oct 2009 – Sept 2010	**Project Coordinator**
	New York Psychiatric Institute, New York, NY
	Research examining suicidality following adverse pregnancy outcomes and research investigating the mental health consequences of genocide and war
Oct 2009-Jan 2010	**Research Assistant,**
	Harlem Children's Zone Evaluation Project, New York, NY
	Research study evaluating effectiveness of an educational program for children in low-income and at-risk populations.
Sept 2008- Dec 2009	**Research Assistant**
	New York Psychiatric Institute, Child Psychiatry, New York, NY
	Study examining the psychological effects of the World Trade Center attacks on children and families with high exposure.
Nov 2008- Nov 2009	**Research Assistant**

	Emotion Regulation Lab, Hunter College, New York, NY *Study using electroencephalogram (EEG) technology to examine emotion regulation in children and adults.*
June 2005-July 2007	**Head Inclusion Teacher** *A.C.E. Integration Head Start*, Brooklyn, N.Y. *Planned, implemented and adjusted curriculum in an integrated classroom designed to meet the educational needs of a diverse group of children with learning, developmental and emotional delays.*
July 2004- Feb 2005	**Assistant House Director** *Rediscovery House, Inc*., Waltham, MA *Oversaw the educational and professional components of a group home for teenage males in state custody to help in their transition to independent living situations.*

Teaching/Invited lectures

2011	**Guest Lecturer**: *Mental Health Measurement in the Context of Conflict and Displacement*, Prof. Courtland Robinson, Johns Hopkins Bloomberg School of Public Health.
2012	**Guest Lecturer**: *Mental Health and Human Rights,* Prof. William Davis, Johns Hopkins University
2013	**Teaching Assistant**: *Issues in Global Mental Health Research*, Prof. Judith Bass, Johns Hopkins Bloomberg School of Public Health.
2014	**Invited Speaker:** *Commonalities in cross-cultural symptomatology for depression and posttraumatic stress disorder,* United States Agency for International Development
2014	**Invited Speaker:** *Implementation science in low-resource settings: Implementation of an evidence-based psychotherapy program in Myanmar,* Washington University in St. Louis
2015	**Teaching Assistant:** *Promoting Mental Health and Preventing Mental Disorders in Developing Countries*, Prof. Wietse Tol, Johns Hopkins Bloomberg School of Public Health.

Languages

Spanish, proficient spoken and written language